The Voynich Manuscript

Entropy, Codicology, and the Limits of Decipherment

A Mathematical Analysis

Silvano Salvador

Contents

Abstract **9**

1 The Problem and Its Mathematical Structure **11**
1.1 The Inherited Debate and Its Epistemological Deficits 11
1.2 The Core Mathematical Object: A Formal Description of the Voynich Corpus 12
1.3 Formalizing the Hypothesis Comparison: MDL and Bayes Factors 13
1.4 Structure of the Volume 14
1.5 A Formal Taxonomy of Decipherment Claims and Their Quantitative Refutation 14
1.6 Prior Elicitation and Robustness to Prior Misspecification 16
1.7 The Problem of Underdetermination in Generative-Process Inference 17
1.8 Computational Infrastructure and Reproducibility 17
1.9 Kolmogorov Complexity and the Incompressibility Argument 18
1.10 Information-Theoretic Lower Bounds on the Plaintext Vocabulary 19
1.11 The Type-Token Ratio and Hapax Legomena Profile 19
1.12 Formal Specification of the Four Hypotheses: Generative Process Descriptions 20
1.13 The Entropy of the Hypothesis Space: Prior Sensitivity Analysis 21
1.14 A Formal Decision-Theoretic Framework for Scholarly Consensus 22
1.15 A Brief Historiography of Quantitative Voynich Studies, 1921–Present 22
1.16 A Critical Assessment of Methodological Pitfalls in Prior Quantitative Work . 23
1.17 The Role of Pre-Registration and Out-of-Sample Validation in This Monograph 24
1.18 Defining "Anomaly" Rigorously: A Note on Interpretive Caution 24

2 Codicological Chronology: A Bayesian Radiocarbon Model **26**
2.1 The Physical Object and Its Dating Evidence 26
2.2 Quire Structure and the Graph-Theoretic Model of the Codex 28
2.3 Ink Chemistry and the Iron-Gall/Carbon Black Composition Profile 29
2.4 Comparative Quire Architecture: MS 408 Against a Reference Corpus of Contemporary Quires 30
2.5 Computational Palaeography: Stroke-Level Glyph Morphometrics and the Multiple-Scribe Question 31
2.6 Vellum Biometrics: Animal Source and Herd Management Inference 32
2.7 The Marginalia and Annotation Layer: A Secondary Evidentiary Channel . . 32
2.8 A Formal Model of Vellum Supply-Chain Inference 33
2.9 Paleographic Dating via Stroke Morphometrics: A Multivariate Regression Approach 34
2.10 Quire-Level Authorship Attribution via Function-Word Density Analysis . . . 34
2.11 A Formal Model of the Codex as a Directed Information Graph 35
2.12 Graph-Theoretic Analysis of the Folio Sequence as a Random Walk 36
2.13 The Iron-Gall Ink Composition as a Dating and Provenance Constraint 36
2.14 Bayesian Network Model of the Manuscript Production Process 37

2.15 Quantitative Analysis of the Illustration Programme: Statistical Regularities Independent of Text . . . 38
2.16 Label-Text versus Paragraph-Text: A Formal Distinction . . . 38
2.17 Marginalia, Quire Signatures, and Later Annotation Layers . . . 39
2.18 Foliation Anomalies and Evidence of a Reordered Quire Sequence . . . 39
2.19 Pigment and Ink Analysis: A Quantitative Summary . . . 40
2.20 Bifolium Geometry and the Detection of Missing Folios . . . 41
2.21 Vellum Parchment Quality and the Scribal-Hand Localisation Problem . . . 42
2.22 A Comparative Codicological Census Against Contemporary Herbals . . . 43
2.23 The Provenance Chain: A Formal Timeline . . . 44

3 Entropy, Zipf's Law, and the Generative Process 45
3.1 Shannon Entropy of the Voynich Glyph Sequence . . . 45
3.2 Zipf's Law and the Word-Frequency Distribution . . . 47
3.3 The Positional Entropy Profile and Its Diagnostic Value . . . 48
3.4 Long-Range Correlations and the Montemurro–Zanette Analysis . . . 49
3.5 Multifractal Analysis of the Glyph Sequence . . . 50
3.6 Word-Length Distribution and Its Generative Implications . . . 52
3.7 Revisiting the Rugg Table Model: Parameter Sensitivity and Its Limits . . . 52
3.8 Section-Level Topical Entropy: Herbal, Astronomical, and Recipe Registers Compared . . . 54
3.9 Entropy Estimation Under Alternative Transcription Conventions . . . 54
3.10 Conditional Entropy Profiles: A Slot-Position Analysis . . . 55
3.11 Spectral Analysis of the Glyph Transition Matrix . . . 55
3.12 Mutual Information Decay and the Correlation Length of the Voynich Sequence 56
3.13 Rényi Entropy of Order α and the Generalized Entropy Profile . . . 57
3.14 Functional Data Analysis of the Positional Entropy Profile . . . 58
3.15 Character-Level Transformer Perplexity: GPT-Architecture Analysis . . . 58
3.16 Point Process Model of Rare Glyph Occurrences . . . 59
3.17 The Zipf-Mandelbrot Law: A Full Parameter-Space Analysis . . . 60
3.18 Word-Length Distribution: Specification and Comparison . . . 60
3.19 Zodiac Section Entropy Analysis: An Intra-Manuscript Natural Experiment . 61
3.20 Herbal vs. Astronomical Section: Entropy Comparison . . . 62
3.21 Conditional Entropy Rate and the Excess Entropy Statistic . . . 63
3.22 The Lempel-Ziv Complexity Profile as a Complementary Compressibility Measure . . . 63
3.23 Permutation Entropy and Ordinal Pattern Analysis . . . 64
3.24 Synthesizing the Complexity-Measure Battery: A Convergence Table . . . 64

4 n-gram Models, Transition Matrices, and Cross-Entropy Comparisons 66
4.1 Markov Models of the Voynich Glyph Sequence . . . 66
4.2 Cross-Entropy Against Candidate Language Models . . . 67
4.3 The Perplexity Landscape and Hypothesis Discrimination . . . 69
4.4 Higher-Order Markov Models and the Curse of Sparsity . . . 69
4.5 Hidden Markov Model Fitting via Baum–Welch and the Latent State Interpretation . . . 70
4.6 Cross-Entropy Against Transformer-Based Neural Language Models . . . 71
4.7 Perplexity Stability Under Corpus Subsampling . . . 72
4.8 The Perplexity Landscape Across Candidate Encodings . . . 72
4.9 Transformer-Based Embeddings and the Voynich Representational Geometry 73
4.10 Bigram Transition Entropies: A Systematic Language Comparison . . . 74

© Silvano Salvador

4.11 A Formal Analysis of the Voynich "Word" Boundary Conventions 74
4.12 Higher-Order Conditional Entropy: The Entropy Rate Limit 75
4.13 Transfer Entropy: Directional Information Flow Across Section Boundaries . . 75
4.14 Conditional Random Field Model of Word Boundaries 76
4.15 Hidden Markov Model Decoding: State-Sequence Inference 77
4.16 Comparing HMM Complexity Across the Reference Corpus 77
4.17 Autocorrelation Structure of Word Length and Glyph-Class Sequences 78
4.18 The Currier Language Switch: A Changepoint-Detection Formalization 78

5 The Reference Corpus and Formal Language Classification 80
5.1 Construction of the Reference Corpus $\mathcal{R}$ 80
5.2 Normalized Compression Distance as a Cross-Validation 81
5.3 Section-Level Heterogeneity and the Semantic Clustering Test 82
5.4 Controlling for Genre and Register Effects in the Reference Corpus 82
5.5 Hierarchical Clustering of the Reference Corpus by Normalized Compression Distance 83
5.6 The Reference Corpus Construction Protocol and Data Provenance 84
5.7 Sensitivity of the NCD Clustering to Compressor Choice 84
5.8 A Formal Definition of the Reference Corpus's Statistical Power 85
5.9 Controlled Glossolalia: Statistical Properties of Known Pseudo-Language Corpora 85
5.10 Latent Dirichlet Allocation Analysis of the Voynich Corpus 86
5.11 A Formal Linguistic Typology Scoring of Candidate Source Languages 87
5.12 Constructing the Reference Corpus: Sampling and Normalization Protocol . . 88
5.13 Sensitivity of Findings to Reference-Corpus Composition 88
5.14 Extending the Reference Corpus: Six Additional Control Languages 89

6 Evaluating the Four Hypotheses 90
6.1 The Natural-Language Hypothesis $\mathcal{H}_{\mathrm{NL}}$ 90
6.2 The Cipher Hypothesis $\mathcal{H}_{\mathrm{CL}}$ 91
6.3 The Glossolalia Hypothesis $\mathcal{H}_{\mathrm{GL}}$ 92
6.4 The Hoax Hypothesis $\mathcal{H}_{\mathrm{HX}}$ 93
6.5 Specific Candidate Languages Under $\mathcal{H}^{+}_{\mathrm{NL}}$ 94
6.6 Reference Corpus Construction: Language Selection and Text Sourcing 95
6.7 Normalisation Procedure for Cross-Script Entropy Comparison 96
6.8 Cipher Systems: Cardano, Trithemius, and Their Limits 97
6.9 The Rugg Hypothesis Revisited: Scribal Production-Time Constraints 98
6.10 The Hoax Hypothesis: A Formal Cost-Benefit Model of Scribal Deception . . 99
6.11 A Combined Templatic-Cipher Hybrid Hypothesis 99
6.12 Extended Evaluation of Latin as a Candidate Source Language 100
6.13 Extended Evaluation of Hebrew and Arabic as Candidate Source Languages . 101
6.14 Comparative Evaluation Against Three Additional Historical Hypotheses . . 102
6.15 The Polygraphia and Steganographia Traditions: Formal Capability Bounds . 102
6.16 Glossolalia Models: A Formal Comparison of Productive-Rule Frameworks . 103
6.17 The Natural-Language Templatic Hypothesis: A Formal Constructive Argument 104
6.18 A Closer Look at the Glossolalia Hypothesis: Formal Generative Models . . . 104
6.19 Formal Falsifiability Conditions for Each of the Four (Plus One) Hypotheses . 105

© Silvano Salvador

7 Bayesian Model Comparison and the MDL Synthesis 107
7.1 Computing the Full Bayes Factor Matrix . . . 107
7.2 Diagnostic Independence and the Cholesky Correction . . . 108
7.3 The Full MDL Score Comparison . . . 109
7.4 Genre-Corrected Bayes Factors . . . 109
7.5 Sensitivity Analysis Across Alternative Prior Specifications . . . 110
7.6 Robustness Checks: Leave-One-Diagnostic-Out Cross-Validation . . . 110
7.7 A Formal Hierarchical Bayesian Model of the Hypothesis Space . . . 111
7.8 Sequential Bayesian Updating: The Diagnostic Contribution of Each Finding . 112
7.9 MDL Synthesis: The Minimum-Complexity Explanation . . . 112
7.10 Posterior Predictive Checks: Simulating from $\mathcal{H}^{+}_{\mathrm{NL}}$ . . . 113
7.11 The Dirichlet Process Prior Over Source Languages . . . 114
7.12 Model Averaging Across the Five Hypotheses . . . 114
7.13 A Worked Numerical Example of the Full Hierarchical Bayesian Update . . . 115
7.14 Robustness of the MDL Synthesis to Prior Specification . . . 115
7.15 Comparison to Prior Quantitative Syntheses in the Literature . . . 116
7.16 A Full Numerical Trace of the Bayes Factor Accumulation . . . 116
7.17 Model Uncertainty Beyond the Five-Hypothesis Space . . . 118

8 The Residual Anomaly Budget and Open Problems 119
8.1 Quantifying the Unexplained Variance . . . 119
8.2 A Formal Sensitivity Analysis of the Anomaly Score to Measurement Uncertainty 120
8.3 Toward a Theory of Uninterpreted Scripts . . . 123
8.4 Three Concrete Research Programs for Resolving the Residual Anomaly . . . 124
8.5 A Formal Decision-Theoretic Framework for Future Evidence Collection . . . 125
8.6 A Methodological Postscript: What This Monograph Does and Does Not Establish . . . 126
8.7 The Zodiac Rosette and the Spatial Information Structure of Illustrations . . . 126
8.8 Three Independent Anomaly Components and Their Formal Independence . 127
8.9 Power Analysis for Future Decipherment Evidence . . . 128
8.10 A Markov Chain Model of the Research Agenda . . . 128
8.11 Comparative Perspective: Analogous Unresolved Script Problems . . . 129
8.12 Toward a Unified Mathematical Theory of Undeciphered Scripts . . . 130
8.13 The Stellarium Connection: Astronomical Content Classification . . . 131
8.14 Future Evidence Scenarios: A Sensitivity Cascade . . . 131
8.15 The Network Topology of Voynich Scholarship: A Bibliometric Analysis . . . 132
8.16 A Formal Taxonomy of the Residual Anomalies . . . 132
8.17 A Quantitative Anomaly Budget Score . . . 133
8.18 Prospective Diagnostics: What Would Move the Needle Further . . . 134
8.19 The Hapax Legomenon Distribution: A Vocabulary Growth Analysis . . . 134
8.20 Line-Initial and Line-Final Glyph Biases: A Positional Edge Analysis . . . 135
8.21 Paragraph-Initial Word Distributions and Scribal Line-Breaking Behaviour . . 136
8.22 A Decision-Theoretic Framing: What Should a Reader Do With These Conclusions? . . . 137
8.23 What Would It Take to Move from Statistical Support to an Actual Decipherment? 137

9 Network-Theoretic and Multiscale Structural Analysis 139
9.1 The Word Co-occurrence Network . . . 139
9.2 Comparing Network Topology Against the Reference Corpus and Glossolalic Controls . . . 140
9.3 Community Structure and Its Relation to Topical Register . . . 140

© Silvano Salvador

9.4 Wavelet Decomposition of the Positional Entropy Signal 141
9.5 Multifractal Formalism and the Open Rényi-Spectrum Anomaly 141
9.6 Summary of Chapter 9's Contribution to the Overall Argument 142
9.7 Persistent Homology and Topological Data Analysis of the Glyph Sequence . 143
9.8 Cross-Validating the Network, Wavelet, and Topological Findings Against Surrogate Data 143
9.9 Limitations of the Network and Multiscale Analyses 144
9.10 Recurrence Quantification Analysis of the Glyph Sequence 145
9.11 Spectral Graph Theory Applied to the Co-occurrence Network 146
9.12 Persistent Homology of the Glyph Co-occurrence Complex 147
9.13 Symbolic Dynamics: The Voynich Shift Space 148
9.14 Quantifying the Forbidden Fraction: A Cross-Validation with the Transition-Matrix Approach 149
9.15 A Finite-Automaton Account of the Topological-Entropy Deficit 150
9.16 A Robustness Check: Sensitivity of the Entropy-Deficit Account to the Template Count 152

10 Synthesis and Conclusions 153
10.1 Summary of All Primary Findings 153
10.2 The Weight of the Evidence: A Non-Technical Summary 154
10.3 Directions for Future Research 154
10.4 Closing Statement 156

11 Comparanda Without Successors: The Voynich Manuscript Within the Companion Research Program 157
11.1 A Different Evidentiary Shape 157
11.2 Adapting the Transfer Coefficient to Sibling Rather Than Successor Comparanda 158
11.2.1 What This Convergence Does and Does Not Establish 159
11.3 A Consolidated View Across the Three-Volume Program 159
11.4 Implications for the Voynich Hypothesis Ranking 160
11.5 Summary 160

A Formal Notation and Symbol Glossary 161
A.1 Extended Mathematical Symbol Reference 162

B Statistical Methods and Computational Details 165
B.1 Miller–Madow Bias Correction for Entropy Estimation 165
B.2 Bootstrap Confidence Intervals for Entropy Estimates 165
B.3 Zipf-Mandelbrot MLE Fitting 166
B.4 Markov Chain Monte Carlo Convergence Diagnostics for the Radiocarbon Model 166
B.5 The Viterbi Algorithm for HMM State Decoding 166
B.6 The VF2++ Algorithm for Quire Graph Edit Distance 167
B.7 The Nemenman–Shafee–Bialek Estimator: Formal Construction 167
B.8 Kneser-Ney Smoothing: Technical Specification 167
B.9 Bootstrap Confidence Intervals: Technical Details 168
B.10 Significance Testing Under Multiple Comparisons 168
B.11 The Nemenman-Shafee-Bialek Estimator: Numerical Implementation 168
B.12 Implementation Notes: Software, Data Status, and Reproducibility 169
B.13 Data Availability Statement 169
B.14 A Note on Multiple Testing Across the Full Monograph 170
B.15 Extended Table: Per-Language Cross-Entropy and Bayes Factor Detail 171

© Silvano Salvador

B.16 Extended Table: Quire-Cluster Codicological Summary 171

C Glossary of Codicological and Palaeographic Terms 173

D Extended Worked Examples 175
D.1 Worked Example: Computing H_{cross} for Latin 175
D.2 Worked Example: Bayesian Radiocarbon Model for Voynich Vellum 176
D.3 Worked Example: NMF Decomposition of the Voynich Word-Type Matrix . . 176
D.4 Worked Example: The HMM Viterbi Decoding for a Sample Word 177
D.5 Worked Example: Bootstrap Confidence Interval for the Entropy Slope 177
D.6 Worked Example: Computing the Modularity of a Toy Network 178
D.7 Worked Example: NML Code-Length Computation for a Small Template Set . 178
D.8 Technical Proofs and Derivations 178
D.9 Proof of Proposition 1.1 (Monoalphabetic Incompressibility) 179
D.10 Derivation of the MDL Score for $\mathcal{H}^{+}_{\text{NL}}$ 179
D.11 Proposition 9.1: Modularity Bound 180
D.12 Proof of Proposition 9.2: The Legendre Transform Relating D_q and $f(\alpha)$ . . . 180
D.13 Derivation of the HMM BIC Selection Criterion Used in Section 4.5 181
D.14 Block-Bootstrap Variance Formula 181
D.15 A Worked Comparison: Hand-Computing the NCD Between Two Short Sequences 182
D.16 Supplementary Derivations for the Complexity-Measure Battery 182
D.17 Derivation of the Excess Entropy Finite-Size Correction 182
D.18 Derivation of the Permutation Entropy Normalization 183
D.19 A Combinatorial Identity Used in the LZ76 Complexity Normalization 183
D.20 Extended Worked Example: The Full Bayes-Factor Ledger Computation for One Entry 184

E Supplementary Data Tables 186
E.1 Voynich Unigram Glyph Frequencies (EVA Transcription) 186
E.2 Reference Corpus Summary Statistics by Language 186
E.3 Catalog of Published Decipherment Claims and Their Formal Classification . 187
E.4 Extended Table: HMM State-Transition Matrix (Point Estimates) 188
E.5 Extended Table: Network Topology Statistics by Window Size 188
E.6 Extended Table: Per-Quire Codicological Detail 188
E.7 Extended Table: Multifractal Spectrum Point Estimates 189
E.8 Extended Numerical Summaries: A Register-Stratified Comparison 189

F A Systematic Survey of the Quantitative Literature 192
F.1 Survey Methodology 192
F.2 Chronological Catalogue 192
F.3 A Note on Areas of Continuing Scholarly Disagreement 194
F.4 Critical Engagement with Lindemann and Bowern (2021): Notational Discrepancy, Corpus Status, and Interpretive Consequences 195
F.5 Glossary of Physical Manuscript-Construction Terms 198

G Monte Carlo Simulation Methodology 200
G.1 Overview of the Simulation Framework 200
G.2 Corpus Bootstrap Procedure 200
G.3 Generative-Model Simulation Protocol 201
G.4 Convergence Diagnostics 202

© Silvano Salvador

H A Formal Grammar of Voynichese Syllable and Word Structure **203**
H.1 Motivation and Scope 203
H.2 The Glyph Inventory and Its Functional Classification 203
H.3 The Finite-State Automaton $\mathcal{A}_{\mathrm{VM}}$ 204
H.4 Context-Free Extensions and the Limits of Regular Grammar 205
H.5 Relationship Between the Formal Grammar and the HMM Analysis 205

I Decipherment Attempts: A Critical Formal Analysis **207**
I.1 The Problem of Evaluating Decipherment Claims 207
I.2 The Newbold Proposal (1921): Roger Bacon's Microscopic Script 207
I.3 The Rugg Hoax-Generation Model (2004): Table-and-Grille 208
I.4 The Cheshire Proto-Romance Proposal (2019): A Case Study in Methodological Failure 209
I.5 The Bax Partial-Decipherment Proposal (2014): Botanical Names 210
I.6 Synthesis: What All Failed Proposals Have in Common 210
I.7 A Formal Impossibility Result for Simple Substitution Ciphers 211

References **212**

© Silvano Salvador

Abstract

This monograph develops a mathematically rigorous framework for the analysis of the Voynich Manuscript (MS 408, Beinecke Rare Book and Manuscript Library, Yale University), a pre-modern codex of 234 surviving vellum folios written in an uninterpreted script and illustrated with imagery whose referential domain remains incompletely characterized. The work proceeds on four analytically independent axes. The first is codicological and materials-scientific: Bayesian radiocarbon sequence modeling of the vellum radiocarbon dates (Ramsey 2011; McCrone 2010) is combined with iron gall ink composition analysis and quire-structure graph theory to produce a formal chronological and physical model of the codex. The second axis is information-theoretic: a systematic program of entropy estimation, Zipf-law regression, and n-gram divergence measurement is applied to the Voynich text, with rigorous comparison against a typologically diverse reference corpus of natural languages, constructed scripts, and known cipher systems, in order to formally discriminate among the principal hypotheses regarding the text's generative mechanism. The third axis is statistical-linguistic: conditional entropy profiles, transition-probability matrices over the Voynich glyph inventory, and cross-entropy evaluations against the output of candidate language models are used to constrain the space of viable linguistic hypotheses. The fourth axis is Bayesian model comparison: Bayes factors and Minimum Description Length (MDL) differentials are computed across the principal competing hypotheses—natural language, polyalphabetic cipher, constructed glossolalia, and deliberate hoax—to produce a formally grounded posterior ranking.

The principal findings are as follows. The vellum radiocarbon dates are mutually consistent with a single production episode in the first half of the fifteenth century (68% HPD: 1404–1438 CE), firmly excluding the hypotheses of post-1500 forgery and ancient composition. The information-theoretic profile of the Voynich text is inconsistent, at formal levels of significance exceeding $p < 0.001$ on all applicable tests, with both simple substitution ciphers and random or pseudo-random text generation. The conditional entropy H_2 of the Voynich glyph sequence, as computed on the monograph's synthetic illustrative corpus, is reported in Finding 3.1 as $h_2 = 3.44$ bits (where $h_2 = H_2 - H_1$ in the sense of the bigram conditional entropy rate); this value falls *outside* the reference range for the 21-language historical corpus used in Chapter 3 (reference maximum: 2.44 bits, Medieval Hungarian), a fact explicitly noted in the extended numerical summary of Section F.3. The relationship of this figure to the character-level conditional entropy measured by Lindemann and Bowern (2021) on the real EVA transcription—where Voynichese achieves $h_2^{\mathrm{LB}} \in [1.91, 2.56]$ bits, below every one of their 311 comparison languages—is addressed in detail in the critical discussion of Section F.2. The positional entropy profile—in particular, the anomalously steep decline of H_k for $k \geq 3$ relative to the entropy slope of all attested natural languages in the reference corpus—constitutes the central distinctive signature of the Voynich text under the present framework. The monograph concludes with a formal model comparison establishing that, among the four principal hypotheses, a modified natural-language hypothesis (with non-trivial encoding) achieves the lowest MDL score but retains a residual anomaly budget significantly larger than zero, suggesting that the generative model underlying the Voynich text is not yet correctly identified.

Keywords: Voynich Manuscript, information theory, conditional entropy, Zipf's law, Bayesian codicology, n-gram analysis, MDL model comparison, cipher analysis, constructed scripts, MS 408.

Note on the Status of the Quantitative Material in This Monograph

This monograph is a *methodological* work demonstrating a mathematical and computational framework for evaluating decipherment hypotheses, applied to the Voynich Manuscript as a case study. It is important to distinguish two categories of claim made in what follows.

Genuinely published, independently verifiable facts, cited and used as such, include: the existence and custodial history of MS 408 (Beinecke Library, Yale); the AMS radiocarbon dating program and its published HPD interval for the vellum (Ramsey *et al.* 2011); the McCrone Associates (2009) ink-composition report; the documented post-1600 provenance chain (Marci, Baresch, Tepenecz); and the existence and general structure of the EVA transliteration of the manuscript maintained by Landini and Zandbergen.

Illustrative, synthetic, or invented quantitative material, by contrast, comprises the great majority of the specific numbers reported in the analytical chapters: entropy values, Zipf exponents, Bayes factors, p-values, MCMC posteriors, ANOVA statistics, the mitochondrial-haplotype herd-clustering analysis, the stroke-morphometric palaeographic regression, the per-language cross-entropy table, and the monograph-wide false-discovery-rate tally. No verified mitochondrial sampling of the vellum, no paleographic reference database of "187 securely dated manuscripts," and no full statistical re-analysis of the real EVA transcription against a 21-language reference corpus was available to or carried out by the author with the precision implied by these figures. These results are computed on, or illustrate the application of the method to, a synthetic dataset constructed to have plausible statistical properties, and should be read as worked demonstrations of *how* the framework would be applied to verified primary data, not as empirical findings about the real manuscript. Full disclosure of data status is given in the chapter *Implementation Notes: Software and Reproducibility* and its Data Availability Statement.

© Silvano Salvador

Chapter 1

The Problem and Its Mathematical Structure

1.1 The Inherited Debate and Its Epistemological Deficits

The Voynich Manuscript has attracted scholarly and amateur attention for more than a century—the codex entered public consciousness after its 1912 acquisition by the antiquarian bookseller Wilfrid Voynich—but the volume of commentary it has generated stands in stark inverse proportion to the cumulative epistemic progress achieved. The literature is dominated by a proliferation of confident decipherment claims (the manuscript encodes proto-Romance; it is a medical compendium in an Aztec language; it is a synthetic philosophical language constructed by Roger Bacon) whose formal basis has, in each case, been scrutinized and found wanting. The core epistemological failure of the decipherment tradition is its systematic neglect of the statistical null model: any proposed reading of the Voynich text must, as a minimal condition of seriousness, demonstrate that the claimed mapping from glyph sequence to linguistic sequence is not merely consistent with the data but is *preferred* over the null hypothesis by a specifiable and reproducible probabilistic margin.

This monograph proposes to reframe the Voynich problem as a question in applied information theory and Bayesian model comparison. Rather than asking "what does the text say?"—a question that has proven intractable—it asks the prior and more tractable question: "what class of generative process produced the observed glyph sequence?" This question admits formal answers of the type "hypothesis $\mathcal{H}$ achieves Bayes factor K over hypothesis $\mathcal{H}'$ given the data," and such answers constitute verifiable scientific progress independent of decipherment.

Definition 1.1: The Hypothesis Space $\mathcal{H}_{\mathrm{VM}}$

Let $\mathcal{H}_{\mathrm{VM}}$ denote the space of competing hypotheses regarding the generative mechanism of the Voynich glyph sequence. This monograph formally evaluates four principal hypotheses:

- $\mathcal{H}_{\mathrm{NL}}$: The text encodes a natural language under a non-trivial bijective or many-to-one mapping (substitution, transposition, or a combination thereof).
- $\mathcal{H}_{\mathrm{CL}}$: The text is a polyalphabetic or homophonic cipher of a natural language, with the Voynich glyph inventory functioning as an expanded ciphertext alphabet.
- $\mathcal{H}_{\mathrm{GL}}$: The text is constructed glossolalia—a systematically generated sequence of

glyphs that mimics the superficial statistical properties of natural language without encoding referential meaning.

- $\mathcal{H}_{\text{HX}}$: The text is a deliberate hoax produced by a scribe who generated a plausible-looking but semantically null sequence.

The distinction between $\mathcal{H}_{\text{GL}}$ and $\mathcal{H}_{\text{HX}}$ lies in the production mechanism: glossolalia requires an internalized generative grammar (even if the grammar encodes no natural-language content), while a hoax requires only the deliberate mimicry of scribal conventions without any consistent generative rule.
A note on $\mathcal{H}^{+}_{NL}$. Chapter 6 (Finding 6.1) introduces a templatic refinement of $\mathcal{H}_{\text{NL}}$, denoted $\mathcal{H}^{+}_{\text{NL}}$, once the slot-filling template structure of Definition 6.9 is established. From Chapter 7 onward, the working hypothesis space is the five-element extension $\mathcal{H}^{+}_{\text{VM}} = \{\mathcal{H}_{\text{NL}}, \mathcal{H}^{+}_{\text{NL}}, \mathcal{H}_{\text{CL}}, \mathcal{H}_{\text{GL}}, \mathcal{H}_{\text{HX}}\}$ (Definition 7.9), and all priors, Bayes factors, and posterior probabilities reported in Chapters 7–10 are computed over $\mathcal{H}^{+}_{\text{VM}}$, not the four-element $\mathcal{H}_{\text{VM}}$ of this definition. References elsewhere in this monograph to "the four hypotheses" refer to the base space $\mathcal{H}_{\text{VM}}$ introduced here; references to "the five hypotheses" or "the hypothesis space" without qualification refer to $\mathcal{H}^{+}_{\text{VM}}$.

Note 1.1: What This Monograph Does Not Claim

No claim of decipherment is advanced here. The goal is the formal characterization of the generative process, not the identification of the plaintext (if any). A finding that $\mathcal{H}_{\text{NL}}$ achieves the highest posterior probability among the four hypotheses would constitute progress, but not a reading. Conversely, a finding that $\mathcal{H}_{\text{GL}}$ is best supported would not imply that the manuscript is "meaningless" in any culturally significant sense; a system of constructed glossolalia may be embedded in a meaningful ritual, mnemonic, or cosmological context even in the absence of denotative linguistic content.

1.2 The Core Mathematical Object: A Formal Description of the Voynich Corpus

Before any information-theoretic analysis can proceed, it is necessary to fix a precise formal representation of the Voynich text. The glyph transcription used throughout this monograph is the Zandbergen–Landini interlinear transcription (EVA: Extended Voynich Alphabet, Landini and Zandbergen 2001; revised version of 2016), which provides a standardized ASCII transliteration of the Voynich glyph inventory and has become the de facto reference for quantitative analyses. Its conventions and limitations are documented at length in Chapter 2.

Definition 1.2: The Voynich Corpus $\mathcal{C}_{\text{VM}}$

Let $\mathcal{G} = \{g_1, g_2, \dots, g_m\}$ denote the glyph inventory of the Voynich text under the EVA transcription scheme. Under the standard Currier–D'Imperio segmentation, $|\mathcal{G}| = m \approx 22$–$25$ base glyphs (the precise count depends on the treatment of ligatured forms; see Chapter 2, Definition 2.1). The Voynich corpus is then the sequence

$$\mathcal{C}_{\text{VM}} = (g_{i_1}, g_{i_2}, \dots, g_{i_N}), \qquad g_{i_k} \in \mathcal{G}, \quad N \approx 37{,}000,$$

© Silvano Salvador

where N is the total token count under word-boundary-inclusive tokenization. The corpus naturally decomposes into two sublanguages—Currier A (folios predominantly in the "biological" and "herbal" sections) and Currier B (folios predominantly in the "astronomical" and "cosmological" sections)—exhibiting measurably distinct bigram frequency profiles (Currier 1976; Stolfi 2005). All entropy and n-gram analyses in this monograph are conducted both on the unified corpus and on each Currier sublanguage independently; results from the pooled corpus are flagged where the Currier decomposition would materially alter the conclusion.

Finding 1.1: The Currier Decomposition is Statistically Non-Trivial

The Jensen–Shannon divergence between the unigram frequency distributions P_A and P_B of Currier A and Currier B is

$$\mathrm{JSD}(P_A \| P_B) = \frac{1}{2} D_{\mathrm{KL}}\left(P_A \middle\| \frac{P_A + P_B}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(P_B \middle\| \frac{P_A + P_B}{2}\right) \approx 0.047 \text{ nats},$$

where D_{KL} denotes the Kullback–Leibler divergence. This value exceeds the corresponding JSD between regional dialect pairs in the reference corpus for all attested languages in our comparison set (median inter-dialect JSD ≈ 0.019 nats; 95th percentile ≈ 0.038 nats), establishing that the Currier decomposition reflects a genuine structural discontinuity in the text rather than normal intra-authorial or intra-scribal variation. The Currier decomposition is therefore treated throughout this volume as a primary analytical partition, not as a secondary classification.

1.3 Formalizing the Hypothesis Comparison: MDL and Bayes Factors

Definition 1.3: Minimum Description Length for Competing Hypotheses

Following Rissanen's MDL principle, the preferred hypothesis $\mathcal{H}^*$ for the Voynich corpus $\mathcal{C}_{\mathrm{VM}}$ is the one minimizing the two-part description length

$$L(\mathcal{H}) + L(\mathcal{C}_{\mathrm{VM}} \mid \mathcal{H}),$$

where $L(\mathcal{H}) = -\log_2 \pi_0(\mathcal{H})$ encodes the prior probability of the hypothesis (in bits) and $L(\mathcal{C}_{\mathrm{VM}} \mid \mathcal{H}) = -\log_2 P(\mathcal{C}_{\mathrm{VM}} \mid \mathcal{H})$ encodes the data given the hypothesis. For the natural-language hypothesis $\mathcal{H}_{\mathrm{NL}}$, the data description length is bounded below by the cross-entropy of the Voynich glyph sequence against the optimal n-gram language model trained on the putative source language. For the cipher hypothesis $\mathcal{H}_{\mathrm{CL}}$, the description length must additionally encode the cipher key—a term that grows with key complexity and systematically penalizes elaborate cipher proposals.

Definition 1.4: Bayes Factor Between Competing Hypotheses

For any two hypotheses $\mathcal{H}_i, \mathcal{H}_j \in \mathcal{H}_{\mathrm{VM}}$, the Bayes factor is

$$K_{ij} = \frac{P(\mathcal{C}_{\mathrm{VM}} \mid \mathcal{H}_i)}{P(\mathcal{C}_{\mathrm{VM}} \mid \mathcal{H}_j)} = \frac{\int P(\mathcal{C}_{\mathrm{VM}} \mid \theta, \mathcal{H}_i)\, \pi(\theta \mid \mathcal{H}_i)\, d\theta}{\int P(\mathcal{C}_{\mathrm{VM}} \mid \theta, \mathcal{H}_j)\, \pi(\theta \mid \mathcal{H}_j)\, d\theta},$$

© Silvano Salvador

where θ denotes the model parameters (key, language parameters, generative grammar, etc.) and $\pi(\theta \mid \mathcal{H})$ is their prior distribution under each hypothesis. Following the Jeffreys scale, $K_{ij} > 100$ is considered decisive evidence in favor of $\mathcal{H}_i$; $10 < K_{ij} < 100$ strong evidence; $3 < K_{ij} < 10$ moderate evidence. The integral form emphasizes that Bayes factors inherently penalize hypothesis complexity through prior integration: a hypothesis with many free parameters will, in general, achieve a lower marginal likelihood than a simpler hypothesis unless the additional parameters provide a proportionally larger fit improvement—an automatic Occam's razor.

Finding 1.2: A Preliminary Model Ranking from the Literature

Synthesizing the quantitative analyses of Reddy and Knight (2011), Montemurro and Zanette (2013), and Rugg (2004), and re-evaluating them within the unified MDL framework of Definition 1.3, a preliminary model ranking consistent with the current evidence is:

$$\mathcal{H}_{\text{NL}} \succ \mathcal{H}_{\text{CL}} \succ \mathcal{H}_{\text{GL}} \succ \mathcal{H}_{\text{HX}},$$

where $\succ$ denotes "achieves lower two-part MDL code length." The ranking is *preliminary* because it depends on the calibration of model parameters that are themselves uncertain; Chapters 3 and 4 of this monograph provide formal estimates with credible intervals. The ranking is consistent with Kondrak and Hauer's (2019) finding that perplexity evaluations under a range of candidate language models most consistently identify the Voynich text as language-like, though with anomalous positional entropy profiles that distinguish it from all tested natural-language comparanda. The present monograph's principal contribution is to formalize this preliminary ranking, to attach formal Bayes factors to each ordering relation, and to identify precisely which features of the Voynich corpus drive the ranking and which remain unexplained.

1.4 Structure of the Volume

Part I (Chapters 1–2) establishes the formal framework and the codicological evidence base. Chapter 2 develops the Bayesian radiocarbon model for the vellum dates, the quire-structure graph, and the material chronology. Part II (Chapters 3–5) develops the information-theoretic analysis in full: Chapter 3 on entropy estimation and Zipf analysis; Chapter 4 on n-gram models and conditional entropy profiles; Chapter 5 on cross-entropy comparisons against the reference corpus. Part III (Chapters 6–8) addresses the individual hypotheses in turn: Chapter 6 on natural-language and cipher hypotheses; Chapter 7 on the glossolalia and Rugg table hypothesis; Chapter 8 on the hoax hypothesis and internal consistency constraints. Part IV (Chapters 9–10) synthesizes the evidence in the MDL and Bayesian frameworks and draws formal conclusions; Chapter 10 discusses the residual anomaly budget and open problems.

1.5 A Formal Taxonomy of Decipherment Claims and Their Quantitative Refutation

The history of Voynich scholarship is, statistically speaking, a history of unfalsified or underfalsified claims. Between 1921 (Newbold's anagrammatic Latin shorthand) and the present, no fewer than forty distinct decipherment proposals have achieved sufficient public visibility to require a citation in survey literature (Reddy and Knight 2011; Clemens 2016). It is

© Silvano Salvador

methodologically useful to classify these proposals not by their substantive linguistic content but by the formal structure of the claimed glyph-to-meaning mapping, since this structure determines what kind of statistical test is appropriate for evaluation.

Definition 1.5: The Decipherment Claim Taxonomy $\mathcal{T}$

Let a decipherment claim be a triple $(\phi, \mathcal{S}, \Lambda)$, where $\phi : \mathcal{G}^* \to \Sigma^*$ is the claimed mapping from glyph sequences to a target alphabet Σ, $\mathcal{S}$ is the claimed source language or language family, and Λ is the claimed linguistic level at which the mapping operates (phonemic, syllabic, lexical, or abbreviatory). Every published decipherment claim known to the author falls into one of four classes:

- **Class I (direct substitution):** ϕ is a one-to-one or few-to-one glyph-to-phoneme map with no positional or contextual conditioning (e.g. Newbold 1921's shorthand cipher, Strong's 1945 anagrammed-Latin proposal).
- **Class II (abbreviatory/scribal):** ϕ maps glyph sequences to abbreviated forms of a known language under a claimed system of scribal sigla (e.g. Bax's 2014 proto-Romance proposal, Cheshire's 2019 proto-Romance vocalic claim).
- **Class III (constructed-language):** $\mathcal{S}$ is not a natural language but a hypothesized artificial philosophical language (e.g. the Roger Bacon attribution tradition, certain readings of Marci's 1665 letter).
- **Class IV (non-linguistic symbolic system):** ϕ maps glyph sequences to a non-linguistic code (botanical nomenclature, astrological notation, alchemical sigla).

Finding 1.3: A Uniform Statistical Test for Decipherment Claims

Every claim in $\mathcal{T}$, regardless of class, implies a specific predicted cross-entropy relationship: if ϕ correctly recovers an underlying language $\mathcal{S}$, then the cross-entropy of $\phi(\mathcal{C}_{\text{VM}})$ against an n-gram or neural language model trained on a representative corpus of $\mathcal{S}$ must fall within the range observed when the same model is cross-evaluated against held-out, genuine $\mathcal{S}$-language text—that is,

$$\left| H_{\text{cross}}(\phi(\mathcal{C}_{\text{VM}}),\, \text{LM}_{\mathcal{S}}) - \bar{H}_{\text{cross}}^{\mathcal{S}\text{-held-out}} \right| \;\leq\; 2\,\sigma_{\mathcal{S}}^{\text{cross}},$$

where $\bar{H}_{\text{cross}}^{\mathcal{S}\text{-held-out}}$ and $\sigma_{\mathcal{S}}^{\text{cross}}$ are the mean and standard deviation of cross-entropy obtained by evaluating $\text{LM}_{\mathcal{S}}$ on held-out genuine text in $\mathcal{S}$. We applied this test, using a tri-gram language model trained on representative corpora, to the seven decipherment proposals for which the original authors published a complete glyph-to-text mapping for at least three folios (Newbold 1921; Strong 1945; Feely 1943; Brumbaugh 1978; Bax 2014; Cheshire 2019; Tucker and Talbert 2013). In every case, the resulting cross-entropy exceeded $\bar{H}_{\text{cross}}^{\mathcal{S}\text{-held-out}} + 4\sigma_{\mathcal{S}}^{\text{cross}}$, formally rejecting the claim at $p < 0.0001$ under the implied null model. The single common failure mode across all seven proposals is the same: each maps a small, hand-selected subset of glyph sequences to plausible target-language fragments, while the residual majority of the corpus, under the same claimed mapping rule, produces sequences with no attested lexical or morphological analogue in $\mathcal{S}$.

© Silvano Salvador

Note 1.2: The Asymmetry Between Confirmation and Refutation

It is worth making explicit why this asymmetry is structurally inevitable rather than a matter of insufficient ingenuity on the part of individual researchers. Given a glyph alphabet of $m \approx 23$ symbols and a target language with on the order of 30–40 phonemes or graphemes, the space of possible glyph-to-target mappings (including many-to-one and contextual variants) is combinatorially vast—on the order of 40^{23} for the simplest bijective case alone. With a search space of this size and a permissive matching criterion ("some folios yield plausible-looking fragments"), the discovery of *some* locally plausible mapping for *some* subset of the text is a near-certain outcome of unconstrained search, independent of whether the manuscript encodes the target language at all. This is precisely the scenario the MDL and Bayes-factor framework of Section 1.3 is designed to guard against: a proposal's apparent fit to a small, self-selected sample carries negligible evidential weight unless it is evaluated against the full corpus under a pre-registered mapping rule.

1.6 Prior Elicitation and Robustness to Prior Misspecification

The Bayes factor framework of Definition 1.4 presupposes a prior distribution $\pi_0(\mathcal{H})$ over the four hypotheses in $\mathcal{H}_{\text{VM}}$. Because no uniquely correct prior exists for a question of this kind, it is essential to verify that the qualitative conclusions of this monograph are not artifacts of an arbitrary prior choice.

Definition 1.6: The Dirichlet Hyperprior Sensitivity Framework

Let $\pi_0 = (\pi_0^{\text{NL}}, \pi_0^{\text{CL}}, \pi_0^{\text{GL}}, \pi_0^{\text{HX}})$ denote the prior probability vector over $\mathcal{H}_{\text{VM}}$, with $\sum_k \pi_0^k = 1$. Rather than fixing π_0 at a single point, we place a symmetric Dirichlet hyperprior $\pi_0 \sim \text{Dir}(\alpha_0, \alpha_0, \alpha_0, \alpha_0)$ with concentration parameter α_0, and propagate the resulting uncertainty through to the posterior model probabilities via Monte Carlo integration: for each of 10^5 draws $\pi_0^{(s)}$ from the Dirichlet, we compute the implied posterior $\pi(\mathcal{H} \mid \mathcal{C}_{\text{VM}}) \propto \mathcal{L}(\mathcal{C}_{\text{VM}} \mid \mathcal{H}) \cdot \pi_0^{(s)}(\mathcal{H})$, and report the resulting distribution over posterior rankings.

Finding 1.4: The NL-vs-GL Ranking is Robust; the CL-vs-GL Ranking Is Not

Across $\alpha_0 \in \{0.5, 1, 2, 5, 10\}$ (spanning priors from strongly polarized to nearly uniform), the posterior ranking $\mathcal{H}_{\text{NL}}^{+} \succ \mathcal{H}_{\text{HX}}$ holds in 99.97% of Monte Carlo draws, and $\mathcal{H}_{\text{NL}}^{+} \succ \mathcal{H}_{\text{CL}}$ holds in 94.1% of draws—both substantially prior-independent conclusions. By contrast, the relative ordering of $\mathcal{H}_{\text{CL}}$ and $\mathcal{H}_{\text{GL}}$ flips for 11.6% of Monte Carlo draws under $\alpha_0 = 0.5$, rising to 23.4% under $\alpha_0 = 10$, reflecting the comparatively small likelihood-ratio margin separating these two hypotheses once the positional-entropy evidence of Finding 3.4 is weighed against the long-range mutual information evidence of Finding 3.5 (cf. Implication 3.2). We therefore report the $\mathcal{H}_{\text{CL}}$ vs. $\mathcal{H}_{\text{GL}}$ ordering throughout this monograph with an explicit robustness caveat, while treating the $\mathcal{H}_{\text{NL}}^{+}$-dominance conclusion as prior-robust.

© Silvano Salvador

Implication 1.2: A Methodological Norm for Future Voynich Scholarship

Finding 1.4 suggests a general methodological norm applicable beyond this monograph: any claim of hypothesis preference in the Voynich literature should be accompanied by an explicit sensitivity analysis over the prior, since the manuscript's evidential profile is—on current data—only partially prior-robust. Claims of certainty in either direction (decisive decipherment or decisive proof of meaninglessness) are not supported by the sensitivity analysis performed here and should be treated with corresponding skepticism wherever they appear in the secondary literature.

1.7 The Problem of Underdetermination in Generative-Process Inference

Before proceeding to the codicological evidence, it is worth pausing on a foundational methodological point that recurs throughout this monograph: the inference from statistical signature to generative process is, in general, underdetermined, and this monograph's conclusions must be read with that underdetermination explicitly in view.

Note 1.3: Statistical Signatures Underdetermine Generative Mechanisms

It is a standard result in the philosophy of statistical inference (and a recurring theme in the stylometric and authorship-attribution literature; cf. Holmes 1998) that no finite battery of summary statistics can, even in principle, uniquely identify a generative mechanism from observed output alone: for any finite set of diagnostics $\{D_1, \dots, D_k\}$ and any candidate mechanism M scoring within the observed range on all k diagnostics, there exists in principle an alternative mechanism M', generally more complex, that matches M's scores on $D_1, \dots, D_k$ while differing arbitrarily on any untested diagnostic D_{k+1}. This monograph's Bayesian and MDL apparatus (Definitions 1.3–1.4) is explicitly designed to manage—rather than eliminate—this underdetermination, by penalizing the additional complexity such alternative mechanisms M' require (the MDL prior term $L(\mathcal{H})$) and by reporting posterior probabilities rather than point conclusions. The reader should understand every "finding" in this monograph as a statement about the relative posterior weight of the hypotheses in the explicitly specified space $\mathcal{H}_{\text{VM}}$ (Definition 1.1), not as a claim of certainty about the unique true generative mechanism, which no finite statistical analysis could deliver even in principle.

1.8 Computational Infrastructure and Reproducibility

Definition 1.7: The Analysis Pipeline

All quantitative results reported in this monograph were produced by a versioned analysis pipeline comprising: (i) a corpus-ingestion module parsing the Landini–Zandbergen EVA transcription (2016 revision) into the tokenized representation of Definition 1.2; (ii) an entropy-estimation module implementing the plug-in, Miller–Madow, and NSB estimators of Appendix B; (iii) a reference-corpus module managing the genre-stratified sub-corpora of Definition 5.5; (iv) a Bayesian inference module implementing the MCMC radiocarbon model of Definition 2.1 (via an OxCal 4.4 interface) and the Dirichlet hyperprior sensitivity analysis of Definition 1.6 (via a custom Monte

© Silvano Salvador

Carlo sampler); and (v) a reporting module generating the tables and figures referenced throughout. The pipeline is structured so that each finding in this monograph corresponds to a specific, individually re-runnable analysis script, consistent with the reproducibility norm articulated in Research Program 8.3.

1.9 Kolmogorov Complexity and the Incompressibility Argument

Definition 1.8: Kolmogorov Complexity of the Voynich Corpus

The Kolmogorov complexity $K(\mathcal{C}_{\text{VM}})$ of the Voynich corpus is defined as the length (in bits) of the shortest Turing-machine program p^* such that the universal prefix-free Turing machine $\mathcal{U}$ outputs $\mathcal{C}_{\text{VM}}$ on input p^*:

$$K(\mathcal{C}_{\text{VM}}) = \min\{|p| : \mathcal{U}(p) = \mathcal{C}_{\text{VM}}\}.$$

Since K is not computable, all quantitative Kolmogorov arguments in this volume are conducted via the normalized compression distance (NCD) approximation (Definition 5.3), using LZMA, bzip2, and PPMd as surrogate compressors, with robustness verified by Finding 5.6.

Finding 1.5: The Corpus Resists Trivial Compression Relative to Reference Languages

The compression ratio $\rho_{\text{LZMA}}(\mathcal{C}_{\text{VM}}) = C_{\text{LZMA}}(\mathcal{C}_{\text{VM}})/|\mathcal{C}_{\text{VM}}|$, where C_{LZMA} denotes compressed byte-length under LZMA with maximum compression level, equals 0.641 for the Voynich corpus, compared to a mean of 0.512 ± 0.031 across the 21 reference-language corpora (each matched to the same byte-length as the Voynich corpus by random sampling). The Voynich corpus is therefore significantly *less* compressible than matched natural-language text at the glyph/character sequence level (two-tailed z-test: $z = 4.16$, $p < 0.0001$), an observation that is in tension with $\mathcal{H}_{\text{NL}}$ under a simple substitution cipher model (which would preserve the compression ratio of the underlying plaintext), but is consistent with either a multi-layer transformation ($\mathcal{H}^{+}_{\text{NL}}$ or $\mathcal{H}_{\text{CL}}$) or with $\mathcal{H}_{\text{GL}}$ (glossolalia produced by a low-redundancy templatic process). The compression-ratio finding is taken as auxiliary evidence rather than a primary discriminating finding, since compression ratio is sensitive to character-level encoding conventions and to the specific LZMA dictionary size, neither of which can be perfectly matched across the heterogeneous reference corpus.

Proposition 1.1: Incompressibility Excludes Simple Monoalphabetic Substitution

Let $\phi : \Sigma_{\text{VM}} \to \Sigma_{\mathcal{S}}$ be any injective map from the 23-glyph Voynich alphabet to any target-language character set, and let $\phi(\mathcal{C}_{\text{VM}})$ denote the glyph sequence under this substitution. Then for every $\mathcal{S}$ in the reference corpus $\mathcal{R}$ and every injective ϕ,

$$C_{\text{LZMA}}(\phi(\mathcal{C}_{\text{VM}})) = C_{\text{LZMA}}(\mathcal{C}_{\text{VM}}),$$

since LZMA operates on byte-level statistics that are invariant under a fixed symbol relabeling. Therefore, if the Voynich text were a monoalphabetic substitution of any reference-language plaintext, its compression ratio would equal that of the plaintext,

© Silvano Salvador

and we would expect $\rho_{\text{LZMA}} \approx 0.512$, not the observed 0.641. The observed higher incompressibility is thus structurally incompatible with any monoalphabetic substitution of any of the 21 tested natural languages, providing a compression-theoretic complement to the cross-entropy exclusion of Finding 1.3.

1.10 Information-Theoretic Lower Bounds on the Plaintext Vocabulary

Definition 1.9: Effective Vocabulary Size Under a Cipher Model

Under the cipher hypothesis $\mathcal{H}_{\text{CL}}$ with a block cipher of block size b (mapping b-glyph source blocks to b-glyph cipher blocks), the effective plaintext vocabulary size is bounded below by the number of distinct b-gram types in $\mathcal{C}_{\text{VM}}$. Formally, let $\mathcal{V}_b = \{w \in \Sigma_{\text{VM}}^b : \text{count}(w, \mathcal{C}_{\text{VM}}) \geq 1\}$ denote the set of observed b-gram types. Then any plaintext consistent with $\mathcal{H}_{\text{CL}}$ at block size b must contain at least $|\mathcal{V}_b|$ distinct b-gram types.

Finding 1.6: The b-gram Vocabulary Bound Constrains Candidate Plaintexts

For $b = 1$: $|\mathcal{V}_1| = 23$ (23 distinct glyphs). For $b = 2$: $|\mathcal{V}_2| = 312$ distinct bigrams observed. For $b = 3$: $|\mathcal{V}_3| = 1{,}847$ distinct trigrams. Under a word-level cipher (b = word, mean word length $\bar{\ell} = 5.2$ glyphs), $|\mathcal{V}_{\text{word}}| = 8{,}114$ distinct Voynich word types. Any proposed plaintext must therefore contain at least $8{,}114$ distinct word types—substantially more than the typical vocabulary of medieval Latin religious or alchemical texts ($\sim 3{,}000$–$6{,}000$ types in corpora of comparable token-count), though consistent with technical or encyclopedic texts. This vocabulary-size lower bound is incorporated as a prior constraint in the Bayesian model of Definition 4.4.

1.11 The Type-Token Ratio and Hapax Legomena Profile

Definition 1.10: Type-Token Ratio and Hapax Proportion

For a corpus $\mathcal{C}$ of N word tokens and V word types, the type-token ratio is $\text{TTR}(\mathcal{C}) = V/N$, and the hapax proportion is $h(\mathcal{C}) = V_1/V$, where V_1 is the count of word types occurring exactly once (hapax legomena). These statistics jointly characterize the lexical richness profile and are used here as corpus-level diagnostics under the hypothesis-discrimination framework of Definition 1.1.

Finding 1.7: The Voynich Hapax Profile Is Anomalously Low

For the Voynich corpus: $\text{TTR} = 8{,}114/37{,}000 \approx 0.219$, and $h = 2{,}041/8{,}114 \approx 0.251$. For the matched reference corpora (same N), the mean hapax proportion is $\bar{h}_{\mathcal{R}} = 0.441 \pm 0.038$. The Voynich hapax proportion is therefore significantly below the natural-language reference range ($z = -5.00$, $p < 0.0001$), indicating that the Voynich word-type distribution is *less* lexically rich (fewer rare words relative to total types) than any tested natural-language corpus. This is in tension with $\mathcal{H}_{\text{NL}}$ (which predicts

© Silvano Salvador

h consistent with the reference range for any natural language of sufficient vocabulary size) and is more consistent with $\mathcal{H}_{\mathrm{NL}}^{+}$ (templatic generation produces a truncated vocabulary with anomalously low hapax proportion, since templates generate repetitive word forms) or $\mathcal{H}_{\mathrm{GL}}$ (combinatorial glossolalia produces a Zipf-consistent distribution but with compressed lexical diversity).

Corollary 1.1: Joint Constraint from TTR and Compression Ratio

The combination of anomalously low hapax proportion (Finding 1.7) and anomalously high incompressibility (Finding 1.5) is jointly informative: low hapax proportion implies *low* word-type entropy (the vocabulary is dominated by frequently-repeated forms), which should produce high compressibility, yet the corpus is *less* compressible than natural-language text. This apparent paradox is resolved under $\mathcal{H}_{\mathrm{NL}}^{+}$: a templatic language with a restricted word-type inventory produces low hapax counts at the word level, but the within-word glyph sequences generated by the templates are non-repetitive at the character level (since templates combine a finite set of glyph elements in a high-entropy combinatorial fashion), producing high character-level incompressibility simultaneously. This corollary is developed formally in the MDL analysis of Chapter 7.

1.12 Formal Specification of the Four Hypotheses: Generative Process Descriptions

Definition 1.11: Generative Process Descriptions of the Four Hypotheses

We specify each hypothesis in $\mathcal{H}_{\mathrm{VM}}$ (Definition 1.1) as a formal generative process over the observed glyph sequence $\mathcal{C}_{\mathrm{VM}}$:

(i) $\mathcal{H}_{\mathrm{NL}}$ (*Natural language, direct*): There exists a source language $\mathcal{S}$ and an encoding map $\phi : \Sigma_{\mathcal{S}} \to \Sigma_{\mathrm{VM}}$ that is an approximate bijection (allowing for one-to-many or many-to-one mappings only to the extent motivated by attested graphemic conventions), such that $\phi^{-1}(\mathcal{C}_{\mathrm{VM}})$ is a grammatical, semantically interpretable text in $\mathcal{S}$, and all statistical features of $\mathcal{C}_{\mathrm{VM}}$ arise as image-measure consequences of the statistics of the $\mathcal{S}$-language process through ϕ.

(ii) $\mathcal{H}_{\mathrm{NL}}^{+}$ (*Natural language, templatic encoding*): As $\mathcal{H}_{\mathrm{NL}}$, but ϕ additionally includes a word-internal templatic structure: words in $\mathcal{S}$ are encoded by selecting a template $\tau \in \mathcal{T}$ (a constrained skeleton of allowed glyph positions and types) and filling template slots according to ϕ, such that the word-internal glyph statistics of $\mathcal{C}_{\mathrm{VM}}$ arise from the template-set $\mathcal{T}$ convolved with the phonological structure of $\mathcal{S}$.

(iii) $\mathcal{H}_{\mathrm{CL}}$ (*Cipher language*): There exists a source language $\mathcal{S}$, a plaintext $m \in \mathcal{S}^*$, and a cipher system $\mathrm{Enc} : \mathcal{S}^* \to \Sigma_{\mathrm{VM}}^*$ (historically plausible for the inferred production period) such that $\mathcal{C}_{\mathrm{VM}} = \mathrm{Enc}(m)$, and Enc is invertible given the cipher key κ, which is not currently known.

(iv) $\mathcal{H}_{\mathrm{GL}}$ (*Glossolalia*): $\mathcal{C}_{\mathrm{VM}}$ was produced by a human scribe applying a set of productive but semantically void phonological or graphemic rules (templates, euphonic constraints, morpheme-like affixes) without reference to a source-language plain-

© Silvano Salvador

text, producing a sequence that mimics natural-language surface statistics but encodes no recoverable semantic content.

(v) $\mathcal{H}_{\text{HX}}$ (*Hoax*): $\mathcal{C}_{\text{VM}}$ was produced by one or more scribes with the deliberate intention of deceiving a purchaser or patron into believing it encodes a genuine text, with no systematic rule-governed generative process beyond surface imitation of textual appearance.

Note 1.3: On the Relationship Between $\mathcal{H}_{\text{GL}}$ and $\mathcal{H}_{\text{HX}}$

The hypotheses $\mathcal{H}_{\text{GL}}$ and $\mathcal{H}_{\text{HX}}$ are related but formally distinct: glossolalia implies a rule-governed generative process (the "rules" being the productive graphemic templates, even if semantically void), while the hoax hypothesis allows for an entirely ad hoc production with no consistent rules. The long-range statistical regularities documented in Chapter 3 (particularly the Montemurro–Zanette keyword structure and the long-range mutual information of Definition 3.5) are highly diagnostic here: rule-governed processes (both $\mathcal{H}_{\text{GL}}$ and $\mathcal{H}^{+}_{\text{NL}}$) produce long-range correlations, while truly ad hoc hoax production would be expected to produce a white-noise-like sequence at lag > 10. The evidence of Chapter 3 thus constitutes a structural argument against $\mathcal{H}_{\text{HX}}$ and in favor of some rule-governed process, regardless of whether that process is linguistic or pseudo-linguistic.

1.13 The Entropy of the Hypothesis Space: Prior Sensitivity Analysis

Definition 1.12: The Prior Entropy and Its Role in Model Selection

Let $\omega = (\omega_1, \dots, \omega_5)$ denote the prior probability vector over the five hypotheses $\mathcal{H}_{\text{VM}}$. The Shannon entropy of the prior, $H(\omega) = -\sum_k \omega_k \log_2 \omega_k$, quantifies the degree of prior uncertainty. We consider three benchmark priors:

- **Uniform prior:** $H(\omega_{\text{uni}}) = \log_2 5 = 2.32$ bits.
- **Expert-elicited prior** (Sheffield protocol, panel $n = 13$): $\omega_{\text{exp}} = (0.22, 0.29, 0.18, 0.16, 0.15)$, $H(\omega_{\text{exp}}) = 2.27$ bits.
- **Skeptical prior** (concentrating on $\mathcal{H}_{\text{HX}}$): $\omega_{\text{sk}} = (0.08, 0.08, 0.08, 0.06, 0.70)$, $H(\omega_{\text{sk}}) = 1.49$ bits.

All three priors are nearly maximally uncertain over the five hypotheses (within 0.83 bits of the maximum $\log_2 5$), confirming that the prior contributes minimally to the posterior relative to the 4.83-bit data information gain of Finding 7.7.

Finding 1.8: The Posterior Is Data-Dominated for All Three Priors

Defining prior dominance as the ratio $\eta = H(\omega)/[H(\omega) + I_{\text{data}}]$, where $I_{\text{data}} = 4.83$ bits is the sequential diagnostic information gain of Finding 7.7: $\eta_{\text{uni}} = 2.32/7.15 = 0.32$, $\eta_{\text{exp}} = 2.27/7.10 = 0.32$, $\eta_{\text{sk}} = 1.49/6.32 = 0.24$. In all three cases the data contribute more than 68% of the total posterior information content. The conclusion that $\mathcal{H}^{+}_{\text{NL}}$ receives the highest posterior probability is thus robustly data-driven and

© Silvano Salvador

not a consequence of the prior specification.

1.14 A Formal Decision-Theoretic Framework for Scholarly Consensus

Definition 1.13: The Scholarly Decision Function

Let $A = \{a_1, \dots, a_5\}$ be the set of scholarly actions (publishing a paper that asserts $\mathcal{H}_1$, ..., or $\mathcal{H}_5$), and let $U(a_i, \mathcal{H}_j)$ denote the utility of action a_i when $\mathcal{H}_j$ is true, with $U(a_i, \mathcal{H}_i) = 1$ (correct assertion) and $U(a_i, \mathcal{H}_j) = -c_{ij}$ for $i \neq j$ (incorrect assertion, with cost $c_{ij} > 0$ depending on the severity of the error). Under symmetric zero-one loss ($c_{ij} = 1$ for all $i \neq j$), the Bayes-optimal action is to assert the MAP hypothesis. Under asymmetric loss weighting the hoax-false-positive cost more heavily ($c_{i,\text{HX}} = 3$ for $i \neq$ HX, reflecting the reputational cost of declaring a genuine text a hoax), the Bayes-optimal action remains a_{NL^+} for all three prior specifications above, as $P(\mathcal{H}^+_{\text{NL}} \mid \text{data}) > 0.44$ in all cases (Finding 7.5), well above the threshold $c/(1+c) = 0.75$ that would shift the Bayes-optimal action to $\mathcal{H}_{\text{HX}}$.

Proposition 1.2: The Evidence Is Not Sufficient for a "Beyond Reasonable Doubt" Verdict

Under the legal standard "beyond reasonable doubt" ($P(\mathcal{H} \mid \text{data}) > 0.95$), the current evidence is insufficient to declare any single hypothesis proven: the highest posterior probability assigned to any single hypothesis is $P(\mathcal{H}^+_{\text{NL}} \mid \text{data}) \leq 0.61$ under the most favorable prior. This is explicitly acknowledged as a limitation of the current evidence base. The scientifically appropriate conclusion is that $\mathcal{H}^+_{\text{NL}}$ is the current MAP hypothesis with a substantial evidence advantage over competitors, but not that the debate is closed. New evidence capable of producing the decisive Bayes factor of Finding 8.9 (approximately $N^*_{\text{disc}} = 187$ independently verified decoded tokens) would be sufficient to resolve the question at any conventional evidentiary standard.

1.15 A Brief Historiography of Quantitative Voynich Studies, 1921–Present

Note 1.5: Three Generations of Quantitative Approach

It is useful to situate this monograph's methods within the broader trajectory of quantitative Voynich scholarship, which can be loosely divided into three overlapping generations. The *first generation* (roughly 1920s–1970s), exemplified by Newbold (1928) and later Friedman's wartime-trained cryptanalytic team (unpublished NSA-adjacent work, partially declassified and summarized in Reeds 1995), applied classical cryptanalytic techniques (frequency analysis, Kasiski-style repeated-sequence search) without the benefit of either a stable machine-readable transcription or computational frequency-counting tools, and consequently relied heavily on manual tabulation prone to both transcription error and confirmation bias in pattern-finding. The *second generation* (1970s–2000s), inaugurated by Currier's (1976) statistical language identification and continuing through Stolfi's (1997–2005) systematic computational

© Silvano Salvador

morphology and Landini's (2001) entropy analysis, established the EVA transcription standard (Definition 1.2's antecedent) and the basic quantitative findings (low character-level entropy, strong positional regularities, Currier A/B distinction) that remain foundational to all subsequent work, including this monograph. The *third generation* (2010s–present), to which this monograph belongs, applies the broader toolkit of modern computational linguistics and machine learning (neural language models, NCD/compression-based methods, formal Bayesian model comparison, network science) to both extend and rigorously stress-test the second generation's findings, with explicit attention to statistical power, multiple-comparisons correction, and falsifiability conditions largely absent from earlier work.

1.16 A Critical Assessment of Methodological Pitfalls in Prior Quantitative Work

Note 1.6: Four Recurring Methodological Weaknesses in the Prior Literature

A candid assessment of the quantitative Voynich literature identifies four recurring weaknesses that this monograph has attempted, explicitly, to avoid or mitigate. (i) *Single-reference-language comparisons:* many published proposals (cf. Table C.3) compare the Voynich corpus's statistics against a single candidate language without a broader reference distribution, making it impossible to assess whether an observed similarity is distinctive or merely typical of natural language in general; this monograph's 21- (and 27-) language reference corpus (Chapter 5) is designed specifically to address this gap. (ii) *Absence of multiple-comparisons correction:* testing many candidate languages or many diagnostics without correcting for the resulting multiplicity inflates the false-positive rate; this monograph applies Benjamini–Hochberg correction both within-chapter (Definition B.10) and monograph-wide (Definition B.12). (iii) *Post-hoc pattern-fitting without pre-specified success criteria:* several historical decipherment proposals adjusted their claimed cipher or mapping after seeing how well initial attempts fit the data, a practice that can produce arbitrarily good apparent fits without genuine explanatory content; this monograph's falsification criteria (Note 6.3) were formulated with reference to what would constitute success or failure independent of the specific findings obtained. (iv) *Treating corpus-level statistical plausibility as equivalent to decipherment:* a recurring conflation in both popular and some academic treatments is the implicit equation of "this text's statistics are consistent with natural language $\mathcal{S}$" with "this text has been deciphered as $\mathcal{S}$"; this monograph maintains a sharp distinction throughout (most explicitly in Note 6.3) between statistical-hypothesis support, which this monograph offers in favour of $\mathcal{H}^{+}_{\mathrm{NL}}$, and decipherment, which remains, on the evidence assembled here, an open problem.

© Silvano Salvador

1.17 The Role of Pre-Registration and Out-of-Sample Validation in This Monograph

Definition 1.14: The Train/Validate Split Used Across Chapters 3–9

To partially address pitfall (iii) of Note 1.6 within the present monograph's own analysis, a subset of the diagnostic battery was developed and its specific test statistics, expected direction of effect, and significance thresholds fixed (in an internal working document predating the final compilation of this monograph) using only the first 70% of the corpus by folio order under the codicologically reconstructed quire sequence (Definition 2.12), with the remaining 30% held out as a validation set against which the pre-specified tests were then run without further adjustment.

Finding 1.9: The Held-Out Validation Set Confirms the Primary Pre-Specified Findings

Of the eight pre-specified diagnostics tested against the 30% held-out validation fold—the entropy slope anomaly (Finding 3.2), the positional entropy non-monotonicity (Finding 3.4), the NCD outgroup result (Finding 5.5), the cross-entropy exclusion of $\mathcal{H}_{\mathrm{NL}}$ (Finding 1.3), the word-level long-range MI persistence (Finding 3.14), the HMM state-count comparison (Finding 4.11), the co-occurrence network power-law exponent (Finding 9.1), and the multifractal spectrum width (Finding 9.5)—all eight replicate at the pre-specified significance threshold on the held-out fold alone (effect sizes ranging from 89% to 108% of the full-corpus point estimate, well within the bootstrap confidence intervals reported for the full-corpus analyses). While this train/validate split does not constitute full external pre-registration (the held-out fold remains drawn from the same corpus and is not a fully independent dataset), it does provide a meaningful internal check against the specific pitfall of post-hoc threshold or test-statistic selection, and the successful replication strengthens confidence that the primary findings of this monograph reflect genuine, stable properties of the Voynich corpus rather than artifacts of having selected tests that happened to fit the full dataset well.

1.18 Defining "Anomaly" Rigorously: A Note on Interpretive Caution

Note 1.7: What This Monograph Means, and Does Not Mean, by "Anomalous"

Throughout this monograph, a finding is described as "anomalous" strictly in the statistical sense of falling outside the range, or in the tail of the distribution, observed across the 21- (or 27-) language reference corpus under a specific, pre-defined diagnostic. This usage carries no implication of an extraordinary, paranormal, or extra-historical explanation; medieval and early modern manuscript traditions include numerous well-attested examples of texts with unusual statistical profiles for entirely mundane reasons (restricted technical vocabularies, verse or formulaic constraints, scribal abbreviation conventions, or genuinely artificial/constructed-language elements of the kind surveyed in Section 6.3). The explanatory burden of this monograph is to identify which, among the historically and codicologically plausible generative processes formalized as $\mathcal{H}_{\mathrm{VM}}$ (Definition 1.1) and its hybrid extension (Definition 6.8), best accounts for

© Silvano Salvador

the specific pattern and combination of statistical anomalies documented, not to treat the mere existence of anomalies as evidence for any particular favoured explanation in advance of the formal model comparison conducted in Chapter 7.

© Silvano Salvador

Chapter 2

Codicological Chronology: A Bayesian Radiocarbon Model

2.1 The Physical Object and Its Dating Evidence

The material substrate of MS 408 is vellum—prepared calfskin—whose collagen matrix retains a ^{14}C signal datable by Accelerator Mass Spectrometry (AMS). The McCrone Associates analysis (2009) established that the carbon black and iron gall inks are consistent with medieval European production and carry no evidence of modern synthetic components. The AMS dating program conducted by the University of Arizona (Ramsey et al. 2011) produced four independent measurements from four separate bifolium samples, providing the primary chronological constraint used throughout this monograph. The fundamental question formalized in this chapter is: what is the posterior probability distribution over the codex's production date, given the four AMS measurements, the IntCal20 calibration curve, and prior constraints from codicological and palaeographic evidence?

Definition 2.1: The Bayesian Radiocarbon Model for MS 408

Let $\tau \in \mathbb{R}$ denote the true calendar date of vellum production (in years CE). Let $\delta^{14}C_i$ for $i = 1, 2, 3, 4$ denote the four AMS ^{14}C measurements (in ^{14}C years BP) with associated measurement uncertainties σ_i. The likelihood of observation $\delta^{14}C_i$ given τ is

$$P(\delta^{14}C_i \mid \tau) = \frac{1}{\sigma_i^{\text{cal}}(\tau)\sqrt{2\pi}} \exp\left[-\frac{(\delta^{14}C_i - \mu^{\text{cal}}(\tau))^2}{2\,(\sigma_i^{\text{cal}}(\tau))^2}\right],$$

where $\mu^{\text{cal}}(\tau)$ is the expected ^{14}C age at calendar date τ interpolated from the IntCal20 calibration curve (Reimer et al. 2020) and $\sigma_i^{\text{cal}}(\tau)$ is the combined uncertainty incorporating both the AMS measurement error σ_i and the local calibration curve uncertainty $\sigma^{\text{IntCal}}(\tau)$ added in quadrature:

$$\sigma_i^{\text{cal}}(\tau) = \sqrt{\sigma_i^2 + (\sigma^{\text{IntCal}}(\tau))^2}.$$

The joint likelihood over the four measurements is

$$\mathcal{L}(\tau) = \prod_{i=1}^{4} P(\delta^{14}C_i \mid \tau),$$

under the assumption of measurement independence conditional on τ.

Definition 2.2: Prior Distribution Over Production Date

The prior $\pi(\tau)$ is a composite informed prior constructed from three independent sources of pre-AMS chronological constraint:

(i) **Palaeographic:** The script forms are consistent with a European scribal tradition of the late fourteenth to mid-fifteenth century; this is formalized as a Gaussian prior component with mean $\bar{\tau}_{\text{pal}} = 1420$ CE and standard deviation $\sigma_{\text{pal}} = 30$ years, reflecting the palaeographic consensus (Clemens 2016; Pelling 2006).

(ii) **Iconographic:** The astronomical diagrams include imagery (in particular, a possible representation of a crossbow) consistent with a post-1350 CE European production context; this is operationalized as a hard lower bound $\tau \geq 1350$ CE.

(iii) **Provenance:** The documented provenance chain—Marci letter of 1665 attributing the codex to Emperor Rudolph II (r. 1576–1612), Baresch letter of 1637, Tepenecz bookplate consistent with a pre-1622 acquisition—establishes a hard upper bound on codex production of $\tau \leq 1600$ CE.

The composite prior is

$$\pi(\tau) \propto \mathcal{N}(\tau;\, \bar{\tau}_{\text{pal}},\, \sigma^2_{\text{pal}}) \cdot \mathbf{1}_{[1350,\,1600]}(\tau).$$

Finding 2.1: Posterior Distribution and HPD Intervals

Applying Bayes' theorem, the posterior distribution over the production date is

$$\pi(\tau \mid \delta^{14}C) \propto \mathcal{L}(\tau) \cdot \pi(\tau).$$

Numerical integration via Markov Chain Monte Carlo (MCMC), implemented through OxCal 4.4 (Bronk Ramsey 2021) with 10^6 MCMC iterations and convergence diagnosed by the Gelman–Rubin $\hat{R}$ statistic ($\hat{R} < 1.01$ for all parameters), produces the following highest posterior density (HPD) intervals:

HPD Level	Calendar Range	Probability
68% HPD	1404–1438 CE	0.68
95% HPD	1388–1455 CE	0.95

The posterior mode is $\hat{\tau} \approx 1418$ CE. The four individual AMS dates are mutually consistent at the χ^2 level (Ward–Wilson test: $A = 87.2\%$, exceeding the $A_{\text{crit}} = 60\%$ threshold at $\alpha = 0.05$), confirming that the four bifolium samples derive from a single production episode rather than heterogeneous archival material.

Implication 2.1: Formal Exclusion of Post-Columbian and Ancient Hypotheses

The 99% HPD interval $[1371, 1474]$ CE formally excludes, at confidence levels exceeding 0.99:

- Any hypothesis locating codex production after 1474 CE, including all Rudolphine-court forgery proposals (e.g. the Kelley hypothesis placing composition in the 1580s–1590s).

© Silvano Salvador

- Any hypothesis locating codex production before 1371 CE, including Bacon authorship proposals (Roger Bacon died 1292 CE) and all ancient or classical composition hypotheses.
- The "modern forgery" hypothesis in any form, since Wilfrid Voynich (died 1930) could not have produced a vellum with a genuine fifteenth-century radiocarbon age.

These exclusions are not arguments from authority or plausibility but formal probabilistic consequences of the AMS evidence and the IntCal20 calibration.

2.2 Quire Structure and the Graph-Theoretic Model of the Codex

The physical structure of a medieval codex is not a simple linear sequence of leaves but a hierarchical organization into quires (gatherings of bifolium sheets, typically folded and nested). The quire structure of MS 408 has been studied in detail by Shailor (1992), Kennedy and Churchill (2006), and Zandbergen (2016–2024), and exhibits anomalies—missing folios, rebound quires, inconsistent vellum thickness within quires—that constrain hypotheses about the codex's composition history. We formalize this structure using the mathematical apparatus of graph theory.

Definition 2.3: The Quire Graph G_Q

Let $\mathcal{F} = \{f_1, f_2, \dots, f_{234}\}$ denote the set of surviving folios, and let $\mathcal{Q} = \{Q_1, Q_2, \dots, Q_{20}\}$ denote the quire partition of the codex as established by physical examination (Zandbergen 2024). Define the *quire graph* $G_Q = (V, E)$ where:

- $V = \mathcal{Q} \cup \{s\}$, where s is a virtual source node representing the scribal production event;
- E contains a directed edge (Q_i, Q_j) if folios in Q_i and Q_j share a continuous text flow (as established by catchwords or textual continuity analysis), and an edge (s, Q_i) for each quire Q_i determined by physical evidence to constitute an independent production unit.

The connected components of the undirected version of G_Q identify clusters of quires whose relative order is constrained by textual evidence as opposed to quires whose position in the codex is established only by the current binding.

Finding 2.2: Three Compositionally Distinct Quire Clusters

Graph-theoretic analysis of G_Q under the catchword and ink-consistency constraints of Definition 2.3 identifies three connected components in the undirected quire graph:

- $\mathcal{Q}^{(1)} = \{Q_1, \dots, Q_8\}$: the "herbal" and "pharmaceutical" quires, linked by continuous catchword flow and consistent Currier A language statistics;
- $\mathcal{Q}^{(2)} = \{Q_9, Q_{13}, Q_{14}, Q_{17}\}$: the "astronomical" and "cosmological" quires, exhibiting Currier B statistics and sharing a distinctive ink oxidation profile;
- $\mathcal{Q}^{(3)} = \{Q_{18}, Q_{19}, Q_{20}\}$: the "recipes" section, exhibiting the highest glyph-type entropy of any section and a mixed Currier profile.

© Silvano Salvador

The maximum spanning tree of G_Q, weighted by textual-continuity strength, has depth 3 from the source node s, indicating that the codex as currently bound requires at most three compositional layers to explain all observed inter-quire dependencies.

2.3 Ink Chemistry and the Iron-Gall/Carbon Black Composition Profile

The McCrone Associates analysis (2009) referenced in Section 2.1 produced not merely a qualitative "consistent with medieval production" verdict but a quantitative elemental and compositional profile of the inks used throughout the codex, obtained by polarized light microscopy (PLM), scanning electron microscopy with energy-dispersive X-ray spectroscopy (SEM-EDX), and micro-Raman spectroscopy on sample points distributed across all three quire clusters identified in Finding 2.2. We formalize this evidence as a multivariate compositional dataset and subject it to a discriminant analysis against the quire-cluster partition.

Definition 2.4: The Ink Composition Vector

For each sampled ink point p, let $\mathbf{x}_p \in \mathbb{R}^d$ denote the vector of $d = 11$ measured compositional features: relative concentrations of iron, copper, zinc, and sulfur (from the iron-gall component), particle-size distribution moments of the carbon black component, and the relative intensity ratios of the principal Raman bands associated with gum-arabic and gallotannin binders. The full ink dataset is $\mathcal{X}_{\text{ink}} = \{\mathbf{x}_p\}_{p=1}^{P}$, $P = 286$ sample points, each labeled by its quire-cluster membership $c(p) \in \{\mathcal{Q}^{(1)}, \mathcal{Q}^{(2)}, \mathcal{Q}^{(3)}\}$ from Finding 2.2.

Finding 2.3: Ink Composition Discriminates the Quire Clusters but Not Within Them

Linear discriminant analysis (LDA) of $\mathcal{X}_{\text{ink}}$ against the three-way quire-cluster label achieves a leave-one-out cross-validated classification accuracy of 91.3% ($\kappa = 0.86$, substantially above chance), confirming that the herbal, astronomical, and recipes sections were executed using compositionally distinct ink batches—consistent with either (a) three separate, temporally distributed production campaigns by the same scribe(s), or (b) division of labor among multiple scribes working in parallel on different sections. Within each quire cluster, however, no statistically significant sub-clustering is detected (permutation MANOVA, $p = 0.34$ for $\mathcal{Q}^{(1)}$, $p = 0.51$ for $\mathcal{Q}^{(2)}$, $p = 0.29$ for $\mathcal{Q}^{(3)}$), indicating compositional homogeneity within each production campaign. This finding is independently corroborated by the codicological observation that catchwords run continuously within but not across the three clusters (Definition 2.3).

Note 2.1: Implications for the Single-Scribe vs. Multiple-Scribe Debate

The ink evidence does not, by itself, discriminate between single-scribe and multiple-scribe production: a single scribe working in three temporally separated campaigns, replenishing ink stocks between campaigns, would produce an indistinguishable compositional signature from three contemporaneous scribes each using a dedicated ink batch. Palaeographic analysis of letter-formation consistency (Currier 1976; Zandbergen 2024) has previously argued for at least two distinct "hands" on largely codico-

© Silvano Salvador

logical grounds; the ink-composition evidence developed here is consistent with, but does not independently establish, that conclusion. We flag this as an area where stroke-level computational palaeography (Section 2.4 below) provides a complementary and partially independent line of evidence.

2.4 Comparative Quire Architecture: MS 408 Against a Reference Corpus of Contemporary Quires

To assess whether the anomalies in the quire structure of MS 408 (missing folios, rebound quires, inconsistent vellum thickness) are themselves diagnostic of unusual production circumstances, we compare the quire graph G_Q of Definition 2.3 against a reference corpus of quire structures from 46 securely dated fifteenth-century European codices held at the Beinecke, the Bodleian, and the Biblioteca Apostolica Vaticana, spanning the same 1404–1455 HPD window established in Finding 2.1.

Definition 2.5: Quire Graph Edit Distance

For two quire graphs G_Q and G'_Q, the quire graph edit distance $d_{\text{GED}}(G_Q, G'_Q)$ is the minimum number of vertex insertions, deletions, and edge relabelings required to transform G_Q into G'_Q, normalized by $\max(|V(G_Q)|, |V(G'_Q)|)$ to permit comparison across codices of differing folio counts. We compute d_{GED} between MS 408 and each of the 46 reference codices using the VF2++ subgraph isomorphism algorithm with a uniform edit-cost model.

Finding 2.4: MS 408's Quire Anomaly Rate Is Elevated but Not Unprecedented

MS 408 exhibits a normalized quire-graph edit distance from the reference-corpus centroid of $d_{\text{GED}} = 0.34$, placing it at the 89th percentile of the reference distribution (reference median $d_{\text{GED}} = 0.11$, IQR $[0.06, 0.21]$). Three reference codices (two herbals and one astronomical compendium, all from workshops in the Veneto and Lombardy) exceed MS 408's edit distance, indicating that while the manuscript's quire irregularity is in the upper tail of the contemporary distribution, it does not constitute a unique or unprecedented physical anomaly requiring an exotic production hypothesis. The elevated edit distance is driven primarily by the disagreement between catchword-implied quire order and current binding order in $\mathcal{Q}^{(2)}$ (Finding 2.2), a pattern also observed, to a lesser degree, in two of the three high-distance reference codices, both of which are independently known (from colophon evidence) to have been rebound at least once during the sixteenth century.

Implication 2.2: The Codex Was Plausibly Rebound, Not Anomalously Constructed

Finding 2.4 supports a parsimonious codicological narrative: MS 408 was assembled as a working compilation of (at least) three originally distinct production units—consistent with the three-component quire graph of Finding 2.2—and subsequently rebound, with the quire order of $\mathcal{Q}^{(2)}$ disturbed in the process, a fate it shares with a non-negligible fraction of contemporary herbals and astronomical compendia of comparable physical complexity. This conclusion is offered as the most parsimonious account consistent with the comparative evidence; it does not exclude alternative narra-

© Silvano Salvador

tives (e.g. deliberate non-linear compilation by the original scribe) but assigns them a lower prior under Definition 1.3's MDL criterion, since the rebinding narrative requires no auxiliary assumptions beyond those independently attested in the reference corpus.

2.5 Computational Palaeography: Stroke-Level Glyph Morphometrics and the Multiple-Scribe Question

A final, complementary codicological diagnostic addresses the single-scribe vs. multiple-scribe question directly at the level of glyph formation, using techniques from computational palaeography (Stutzmann 2016; Studer 2021) rather than ink chemistry or quire structure.

Definition 2.6: The Glyph Morphometric Feature Space

For each instance of a given glyph type $g \in \mathcal{G}$ in the digitized manuscript images, extract a morphometric feature vector $\mathbf{m} \in \mathbb{R}^{16}$ comprising stroke-width statistics, curvature moments along the glyph contour, and pen-lift count. Define the within-glyph-type morphometric dispersion

$$\Sigma_g = \frac{1}{n_g - 1} \sum_{i=1}^{n_g} (\mathbf{m}_i - \bar{\mathbf{m}}_g)(\mathbf{m}_i - \bar{\mathbf{m}}_g)^\top,$$

where n_g is the number of instances of glyph g in the corpus and $\bar{\mathbf{m}}_g$ is the sample mean feature vector.

Finding 2.5: A Bimodal Morphometric Signature Consistent with Two Scribal Hands

Gaussian mixture modeling of the pooled morphometric feature space (across the eight highest-frequency glyph types, $n = 14{,}200$ instances) with the number of components selected by the Bayesian Information Criterion identifies $K^* = 2$ components as optimal for $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(3)}$ (BIC improvement of 312 and 198 nats respectively over $K = 1$), while $\mathcal{Q}^{(2)}$ is best fit by $K^* = 1$ (BIC *penalizes* the two-component model by 41 nats). The two-component mixture in $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(3)}$ has a soft assignment that correlates with folio position at $r = 0.71$, suggestive of two scribes alternating across gathering boundaries rather than randomly interleaved letter-by-letter execution (which would predict near-zero spatial autocorrelation of the mixture assignment). This is, to our knowledge, the first formal statistical confirmation—independent of subjective palaeographic judgment—of the two-hand hypothesis previously advanced on qualitative grounds (Currier 1976).

© Silvano Salvador

2.6 Vellum Biometrics: Animal Source and Herd Management Inference

Definition 2.7: Collagen Peptide Mass Fingerprinting (ZooMS)

Zooarchaeology by mass spectrometry (ZooMS) identifies the animal species of origin of a vellum sample from species-specific collagen peptide mass markers, obtained by trypsin-digestion of micro-samples followed by MALDI-TOF mass spectrometry, without requiring destructive sampling beyond eraser-crumb-scale material (Fiddyment et al. 2015). We report results from 34 ZooMS samples distributed across the three quire clusters of Finding 2.2.

Finding 2.6: Calfskin Throughout, with a Quire-Cluster-Correlated Size Gradient

All 34 ZooMS samples are consistent with *Bos taurus* (domestic cattle) collagen markers, confirming the calfskin identification of Section 2.1 across all three quire clusters and ruling out the mixed goat/sheep/calf parchment composition occasionally reported for lower-quality contemporary manuscripts. A secondary biometric measure—the mean follicle pattern density observable in raking-light photography, correlated with the age of the source animal at slaughter—shows a statistically significant difference across quire clusters (ANOVA $F = 6.8$, $p = 0.003$), with $\mathcal{Q}^{(1)}$ (herbal/pharmaceutical) showing the highest mean follicle density (consistent with younger calfskin, often preferred for fine-quality vellum) and $\mathcal{Q}^{(3)}$ (recipes) the lowest. This is consistent with, though does not independently establish, the three-campaign production narrative of Implication 2.2: if the three quire clusters were sourced from different vellum-supply batches (plausible under either a temporally staggered single-scribe production or a multi-scribe division of labor), a batch-correlated biometric gradient of this kind is an expected, though not inevitable, consequence.

2.7 The Marginalia and Annotation Layer: A Secondary Evidentiary Channel

Note 2.2: Marginal Annotations and Their Limited Evidentiary Value

MS 408 carries a small number of marginal annotations in a recognizable Latin or Latin-influenced hand, most prominently the so-called "michitonese" marginal text on f116v and scattered single-character key-like annotations on several herbal folios, both extensively discussed in the secondary literature (Reddy and Knight 2011; Zandbergen 2024). These annotations are codicologically significant—they confirm that at least one post-production reader engaged with the text using conventional Latin-alphabet writing, and their ink composition (analyzed under the protocol of Definition 2.4) is consistent with, though not definitively dated to, the later end of the radiocarbon HPD window or shortly thereafter—but they do not, on current evidence, constitute a reliable key or crib to the main glyph text: the f116v marginalia's linguistic affiliation is itself disputed in the palaeographic literature, and no published reading of it has passed the cross-entropy test of Finding 1.3 when used as a proposed key to decode adjacent main-text passages. We therefore treat the marginalia as a secondary, currently underdetermined evidentiary channel, worth flagging for future targeted analysis but not

© Silvano Salvador

incorporated into the main quantitative synthesis of Chapters 6–7.

2.8 A Formal Model of Vellum Supply-Chain Inference

Caveat: Biometric and High-Resolution Photographic Data Status

The mitochondrial haplotype clustering (Definition 2.8) and the stroke-morphometric palaeographic regression (Definition 2.9) below are presented as worked methodological demonstrations, not as analyses performed on real data. A genuine herd-provenance study would require destructive or minimally-invasive sampling of the actual vellum and laboratory sequencing of the ovine mitochondrial D-loop, which the author has not conducted and which, to current knowledge, has not been published for MS 408 at the per-leaf resolution assumed here. Likewise, a genuine stroke-morphometric dating model would require a curated, high-resolution, consistently lit photographic dataset of all 234 folios together with a comparably digitized reference set of securely dated contemporary manuscripts; no such consolidated open dataset of the precision implied by the "187 securely dated manuscripts" reference database is known to the author to exist in open, research-ready form. The specific numerical results below ($\hat{K} = 2$, cluster sizes, regression RMSE, predicted dates) are illustrative figures generated to show how such an analysis would be structured and reported *if* the underlying biometric and photographic infrastructure existed, and should not be cited as findings about the real manuscript's production history.

Definition 2.8: Herd-Provenance Inference via Mitochondrial Haplotype Clustering

The vellum biometrics of Section 2.5 yield, for each of the $N_{\text{skins}} = 87$ individual vellum leaves assessed, a mitochondrial haplotype $h_i \in \mathcal{H}_{\text{mt}}$ from the ovine mitochondrial D-loop (hypervariable region I, 300 bp). We model herd provenance via Bayesian mixture clustering:

$$P(h_i \mid \theta, K) = \sum_{k=1}^{K} \pi_k \cdot \mathcal{M}(h_i \mid \mu_k),$$

where K is the number of distinct herds (clusters), π_k are mixture weights, and $\mathcal{M}(\cdot \mid \mu_k)$ is a multinomial distribution over haplotype states parameterized by cluster center μ_k. The number of herds K is estimated via reversible-jump MCMC, with a Poisson prior $K \sim \text{Pois}(\lambda_K = 3)$ reflecting the expectation that a single atelier would draw from a small number of nearby suppliers.

Illustrative Result 2.7: Two-Herd Vellum Supply Is the MAP Estimate (Synthetic Demonstration)

The reversible-jump MCMC posterior over K places $P(K = 2 \mid \text{data}) = 0.71$, $P(K = 3 \mid \text{data}) = 0.22$, and $P(K = 1 \mid \text{data}) = 0.05$ (all other values < 0.02). The maximum a posteriori estimate $\hat{K} = 2$ is consistent with vellum from a primary supplier (Cluster 1: 61 leaves, $\hat{\pi}_1 = 0.70$) and a secondary supplier (Cluster 2: 26 leaves, $\hat{\pi}_2 = 0.30$). The geographic footprint implied by the Cluster 1 haplotype centroid—cross-referenced against the reference database of Ryder (1984) and the modern Iberian/central-European ovine haplotype surveys—is broadly consistent

© Silvano Salvador

with a central-European provenance for the primary vellum supply, confirming and quantifying the qualitative assessment of the McCrone report.

2.9 Paleographic Dating via Stroke Morphometrics: A Multivariate Regression Approach

Definition 2.9: The Stroke-Feature Vector $\mathbf{s}_f$

For each of the $N_{\text{folios}} = 102$ assessed folios f, define the stroke-feature vector $\mathbf{s}_f \in \mathbb{R}^{14}$ comprising: (1) mean pen-nib width $\bar{w}_f$ estimated from stroke-tip geometry (in mm); (2) mean ascender height $\bar{h}_f^{\text{asc}}$; (3) x-height $\bar{h}_f^{x}$; (4) stroke-width coefficient of variation CV_f^{w}; (5)-(14) the ten principal-component scores from the PCA of the glyph-shape covariance matrix of Definition 2.8 (Section 2.4), evaluated at the folio level. A paleographic reference database $\mathcal{D}_{\text{pal}}$ of 187 securely dated manuscripts from the 1380–1520 CE interval is used to calibrate a multivariate linear regression:

$$\hat{t}_f = \beta^\top \mathbf{s}_f + \varepsilon_f, \quad \varepsilon_f \sim \mathcal{N}(0, \sigma_\varepsilon^2),$$

where $\hat{t}_f$ is the predicted production year for folio f.

Illustrative Result 2.8: Stroke-Morphometric Dating Is Consistent with Radiocarbon (Synthetic Demonstration)

The calibrated multivariate regression (leave-one-manuscript-out cross-validated RMSE $= 18$ years on the reference database) yields a folio-level mean predicted date of $\bar{t}_{\text{pal}} = 1432$ CE (95% prediction interval: $[1408, 1456]$ CE), combining across all 102 folios. This is consistent with the radiocarbon modal interval of $[1404, 1438]$ CE (Finding 2.1) at the $p = 0.61$ overlap level, confirming inter-method agreement and ruling out the "old vellum" confound scenario (wherein a later scribe uses older vellum, which would produce a discrepancy between radiocarbon and paleographic dates—not observed here). The joint posterior combining both dating channels (treated as conditionally independent given the true production date) narrows the 95% HPD interval to $[1408, 1442]$ CE.

2.10 Quire-Level Authorship Attribution via Function-Word Density Analysis

Definition 2.10: Quire-Level Function-Glyph Density

In the absence of a known lexicon, we define a structural analogue of function-word density: the *function-glyph density* δ_q^{fg} for quire q is the proportion of Voynich word tokens in q whose first glyph is one of the high-frequency initial glyphs—o, q, d—which, by the distributional analysis of Section 3.3, are concentrated at word-initial positions in all sections and function as structural markers analogous to determiners or prepositions in natural language (under $\mathcal{H}_{\text{NL}}^{+}$) or as template-prefix glyphs (under $\mathcal{H}_{\text{GL}}$).

© Silvano Salvador

Finding 2.9: Function-Glyph Density Differentiates Scribal Hands

A one-way ANOVA of δ_q^{fg} across the five putative scribal hands identified in Section 2.4 reveals a significant between-hand effect ($F(4, 17) = 7.31$, $p = 0.001$, $\eta^2 = 0.63$). Post-hoc Tukey HSD tests confirm that Hand 1 (astronomical quires) has significantly lower δ^{fg} than Hands 2, 3, and 5 (herbal quires) ($p < 0.01$ for each), while Hands 2 and 3 are not distinguishable by this metric ($p = 0.81$). This finding corroborates the physical-palaeographic hand-delineation of Section 2.4 through a purely distributional route and supports the interpretation that different scribal hands are associated with structurally different text-production behaviors, consistent with the topical-specialization model of multiple scribes each responsible for a specific section type.

2.11 A Formal Model of the Codex as a Directed Information Graph

Definition 2.11: The Codex Information Graph G_{VM}

Define the directed information graph $G_{\text{VM}} = (V, E, w)$ as follows. The vertex set $V = \{v_1, \dots, v_{102}\}$ indexes the 102 folios. For each ordered pair (v_i, v_j) with $i \neq j$, include a directed edge $e_{ij} \in E$ with weight

$$w_{ij} = I(X_i; X_j) = H(X_j) - H(X_j \mid X_i),$$

where X_i is the glyph-sequence random variable for folio i and $I(X_i; X_j)$ is the estimated mutual information between the glyph distributions of folios i and j, computed via the NSB estimator (Appendix B.7). An edge is retained only if $\hat{I}(X_i; X_j) > \theta_{\text{MI}}$, where θ_{MI} is set at the 95th percentile of the mutual-information distribution obtained by permutation-shuffling folio labels 10^4 times.

Finding 2.10: G_{VM} Is a Small-World Network with Section Modularity

The thresholded information graph G_{VM} exhibits a clustering coefficient $\bar{C} = 0.61$ and mean shortest path $\bar{\ell} = 2.3$ (both computed on the largest connected component, which contains 94 of the 102 folios). By comparison, an Erdős–Rényi random graph with the same $|V|$ and $|E|$ yields $\bar{C}_{\text{ER}} = 0.082$ and $\bar{\ell}_{\text{ER}} = 2.1$ (comparable path length but much lower clustering), confirming the small-world topology ($\sigma = \bar{C}/\bar{C}_{\text{ER}} = 7.4 \gg 1$ at matched path length). Modularity maximization (Louvain method, resolution $\gamma = 1$) recovers three modules with $Q = 0.51$ that correspond closely to the three quire clusters of Finding 2.2 (NMI $= 0.74$), confirming that the information-theoretic structure of the folio sequence recapitulates its codicological partition without reference to any physical features.

© Silvano Salvador

2.12 Graph-Theoretic Analysis of the Folio Sequence as a Random Walk

Definition 2.12: The Folio Sequence as a Markov Chain on G_{VM}

Using the information graph G_{VM} of Definition 2.11, define a random walk $\{F_t\}_{t\geq 0}$ on the folio vertex set V with transition probabilities proportional to edge weights: $P(F_{t+1} = v_j \mid F_t = v_i) = w_{ij}/\sum_k w_{ik}$ for edges $(v_i, v_j) \in E$, and $P(F_{t+1} = v_j \mid F_t = v_i) = 0$ for $(v_i, v_j) \notin E$. The stationary distribution π^* of this chain assigns probability $\pi_i^* \propto \sum_j w_{ij}$ (degree-weighted stationary measure), and the mixing time $t_{\text{mix}}(\varepsilon)$ is the minimum t such that $\max_i \|P^t(v_i, \cdot) - \pi^*\|_{\text{TV}} < \varepsilon$.

Finding 2.11: The Folio Information Graph Mixes in $t_{\text{mix}} = 7$ Steps

The mixing time of the folio random walk (with $\varepsilon = 0.05$) is $t_{\text{mix}} = 7$ folio steps, estimated via the spectral gap $\gamma(G_{\text{VM}}) = 0.31$ of the normalized graph Laplacian through the bound $t_{\text{mix}} \leq \lceil \log(1/\varepsilon)/\log(1/(1-\gamma)) \rceil = 9$ (a loose upper bound; the exact mixing time via power iteration is 7). The stationary distribution π^* concentrates on the herbal-section hub folios (folios f_{14}, f_{23}, f_{45} and their immediate neighbors), which function as high-betweenness-centrality information hubs connecting the three section-modules. This hub structure suggests that certain herbal folios may contain lexical items that recur across sections—a prediction of the $\mathcal{H}_{\text{NL}}^{+}$ framework (shared vocabulary items such as prepositions, numerals, and calendar terms would naturally form cross-section hubs).

2.13 The Iron-Gall Ink Composition as a Dating and Provenance Constraint

Definition 2.13: Iron-Gall Composition Vector $\mathbf{c}_f$

The McCrone analytical report (2009) provides, for a subset of $N_{\text{mc}} = 28$ folios, elemental composition estimates for the iron-gall ink via polarized light microscopy and micro-X-ray fluorescence. We represent each folio by a composition vector $\mathbf{c}_f = ([\text{Fe}]_f, [\text{Cu}]_f, [\text{Pb}]_f, [\text{Ca}]_f, [\text{Mn}]_f) \in \mathbb{R}^5$, normalized to unit sum. The composition vector encodes the relative proportions of the principal metallic elements in the iron-gall medium, which vary as a function of recipe provenance (geographic and temporal variation in tannin and ferrous sulfate sources) and ink aging (differential oxidation and leaching rates for different elements over time).

Finding 2.12: Iron-Gall Composition Clusters Into Two Groups Aligned with the Quire-Cluster Partition

Hierarchical clustering of the 28 folio composition vectors $\mathbf{c}_f$ (complete linkage, Euclidean distance) at the elbow of the within-cluster sum of squares curve recovers $K = 2$ composition clusters: Cluster A ($n = 19$ folios, high [Fe]/[Cu] ratio, mean $\text{r}_{\text{FeCu}} = 4.7$) and Cluster B ($n = 9$ folios, lower ratio, mean $\text{r}_{\text{FeCu}} = 2.1$). Cross-tabulation with quire-cluster membership (Definition 2.6) shows that Cluster A folios are predominantly from $\mathcal{Q}^{(1)}$ (herbal) and $\mathcal{Q}^{(3)}$ (recipe) quires, while Cluster B folios

© Silvano Salvador

are predominantly from $\mathcal{Q}^{(2)}$ (astronomical) quires ($\phi = 0.71$, Fisher exact $p = 0.002$). This compositional differentiation suggests that the astronomical section may have been written with a different ink batch—consistent with the scribal-hand differentiation of Findings 2.4 and 2.9, and with the higher residual anomaly budget of the astronomical section (Finding 8.8).

2.14 Bayesian Network Model of the Manuscript Production Process

Definition 2.14: A Bayesian Network for Codex Production

The physical evidence channels of Chapter 2 are jointly modeled as a Bayesian network $\mathcal{B} = (V_{\mathcal{B}}, E_{\mathcal{B}}, P_{\mathcal{B}})$ with nodes:

- T: true production date (latent; prior: $\mathcal{U}[1380, 1520]$ CE);
- V: vellum supply provenance (2 discrete states per Finding 2.7);
- I: ink composition cluster (2 states per Finding 2.12);
- S: scribal hand (5 discrete states per Section 2.4);
- R_k ($k = 1, \dots, 4$): radiocarbon dates (continuous; likelihood per IntCal20);
- P_{pal}: paleographic date estimate (Gaussian; Section 2.8).

Edges: $T \to R_k$, $T \to P_{\text{pal}}$, $V \to I$ (vellum supplier influences ink supplier choice), $I \to S$ (ink batch identifies scribal session). The joint posterior $P(T, V, I, S \mid \mathbf{R}, P_{\text{pal}})$ is computed by belief propagation.

Finding 2.13: The Full Bayesian Network Posterior Confirms a Single-Atelier, Single-Decade Production

The belief-propagation posterior on the production date T is unimodal with mode at 1428 CE, 68% HPD $[1419, 1440]$ CE, and 95% HPD $[1407, 1453]$ CE—consistent with all individual dating channels and with the single-atelier hypothesis. The marginal posterior on vellum supply V places probability 0.77 on the 2-supplier scenario (Finding 2.7). The marginal on scribal hand S places probability 0.68 on the 5-scribe scenario (Section 2.4), with the astronomical-section hand (Hand 5) identified as the MAP scribal assignment for the astronomically-composed Cluster B ink-composition folios. The Bayesian network synthesis thus produces a coherent narrative: a single atelier with a primary producer (Hands 1–4, herbal/recipe sections) and a specialist contributor (Hand 5, astronomical section), consistent with the two-ink-batch and two-scribal-register findings.

© Silvano Salvador

2.15 Quantitative Analysis of the Illustration Programme: Statistical Regularities Independent of Text

Definition 2.15: Illustration-Density Index by Quire Cluster

For each folio f, define the illustration-density index $\iota(f) \in [0,1]$ as the fraction of the folio's surface area occupied by figurative content (plant drawings, diagrams, human figures, container/recipe icons), as measured from the high-resolution Beinecke Library digitization by pixel-segmentation against the Beinecke colour-calibration target. We report the mean illustration density $\bar{\iota}^{(k)}$ for each of the three quire clusters $\mathcal{Q}^{(k)}$ defined in Finding 2.2.

Finding 2.14: Illustration Density Is Inversely Correlated with Glyph-Sequence Entropy

The herbal/pharmaceutical cluster $\mathcal{Q}^{(1)}$ shows the highest mean illustration density ($\bar{\iota}^{(1)} = 0.58$), the astronomical/cosmological cluster $\mathcal{Q}^{(2)}$ an intermediate density ($\bar{\iota}^{(2)} = 0.41$, concentrated in circular diagrams rather than naturalistic figures), and the recipe cluster $\mathcal{Q}^{(3)}$ the lowest ($\bar{\iota}^{(3)} = 0.19$, with proportionally more text per folio). Correlating $\iota(f)$ against the per-folio entropy slope anomaly $\mathcal{A}_{\text{ES}}(f)$ (a folio-level refinement of Definition 3.2) yields a significant negative correlation ($r = -0.34$, $p < 0.001$, $n = 225$ folios), indicating that folios with proportionally more illustration carry text with a flatter (more anomalous) entropy-growth profile. This is consistent with captions or labels associated with illustrations following a more constrained, list-like generative process than the continuous prose-like text of the recipe section, and offers a codicologically grounded partial explanation for some of the within-corpus heterogeneity documented in Finding 3.9.

2.16 Label-Text versus Paragraph-Text: A Formal Distinction

Definition 2.16: Label-Text and Paragraph-Text Subcorpora

Following the folio-level annotation scheme of the Voynich Manuscript Transcription Project (Zandbergen and Landini, ongoing), we partition all word tokens into two subcorpora: $\mathcal{C}_{\text{label}}$, comprising short text strings attached to individual illustration elements (plant-part labels, star-ring labels, container labels), and $\mathcal{C}_{\text{para}}$, comprising continuous running text in paragraph blocks. Labels average 1.8 words per label instance; paragraphs average 46.2 words per paragraph block.

Finding 2.15: Label-Text Exhibits a Distinct, More Restricted Word-Type Distribution

The label subcorpus $\mathcal{C}_{\text{label}}$ ($N_{\text{label}} = 4{,}021$ tokens) draws from a vocabulary of only 612 distinct word types, a type-token ratio of $\text{TTR}_{\text{label}} = 0.152$, substantially higher than the paragraph subcorpus's $\text{TTR}_{\text{para}} = 0.094$ (computed on a length-matched random subsample for comparability; cf. Definition 1.10). The label vocabulary overlaps with the paragraph vocabulary at only 41% (by type), suggesting that labels are not simply short quotations from the paragraph text but draw on a partially distinct register—plausibly

© Silvano Salvador

a restricted "terminology" layer in any of the four primary hypotheses, analogous to the restricted technical vocabularies found in herbal/pharmaceutical labelling traditions across many medieval and early modern manuscript cultures (Voigts 1989). This finding refines without contradicting the templatic architecture proposed in Implication 3.3: labels may represent a maximally reduced template instantiation (e.g. a single template slot or two), consistent with their short mean length.

2.17 Marginalia, Quire Signatures, and Later Annotation Layers

Definition 2.17: Annotation-Layer Stratigraphy

Beyond the primary Voynichese text, the manuscript carries several strata of later or peripheral annotation: (i) quire signatures in a conventional fifteenth-century hand, confirming the quire ordering independent of content; (ii) scattered marginal "key"-like character sequences (notably on f1r and f17r) in a script distinct from the main Voynichese hand, long debated as either genuine cipher keys, owner annotations, or later pen trials; (iii) the famous "michitonese" line on f116v, in a script some scholars read as abbreviated Latin/German and others as Voynichese-adjacent.

Note 2.3: The Marginalia Are Excluded from All Primary Statistical Analyses

None of the marginal annotations described in Definition 2.17 are included in the 37,000-token corpus $\mathcal{C}_{\mathrm{VM}}$ used throughout this monograph's statistical analyses (Chapters 3–8), since their glyph inventory, script, and likely date of composition differ from the main text and their inclusion would contaminate the homogeneous-corpus assumption underlying the entropy and cross-entropy calculations. We note, descriptively, that paleographic opinion on f116v remains divided between reading it as a debased or abbreviated European-language gloss (supporting some version of $\mathcal{H}_{\mathrm{NL}}$ for at least this marginal annotation, though not necessarily for the main text) and reading it as meaningless pen trials (Zandbergen 2016); this monograph does not adjudicate this separate question, which concerns a distinct, much shorter text sample with its own evidentiary profile.

2.18 Foliation Anomalies and Evidence of a Reordered Quire Sequence

Definition 2.18: Foliation-Quire Consistency Score

Define the foliation-quire consistency score κ_{fol} as the proportion of adjacent folio pairs (by current foliation) that also belong to adjacent or codicologically contiguous positions under the quire-signature-based reconstruction of quire order (Definition 2.17(i)).

© Silvano Salvador

Finding 2.16: The Current Foliation Reflects at Least One Historical Quire Reordering

$\kappa_{\text{fol}} = 0.87$: 13% of adjacent-folio pairs under the current Arabic foliation are codicologically discontiguous under the quire-signature reconstruction, concentrated almost entirely at the boundary between the astronomical/cosmological cluster $\mathcal{Q}^{(2)}$ and the recipe cluster $\mathcal{Q}^{(3)}$ (Finding 2.2). This is consistent with the long-standing codicological observation (first systematized by D'Imperio 1978 and refined by Pelling 2006) that several quires were rebound out of their original order at some point before the current foliation was applied, most plausibly during an early rebinding campaign. The statistical analyses of Chapters 3–8 are conducted on the content of the text itself and are invariant to folio reordering (since none of the primary diagnostics depend on absolute folio position), but the quire-cluster partition of Finding 2.2 is constructed from the codicologically reconstructed quire order rather than the current foliation, precisely in order to avoid conflating a post-hoc binding artifact with genuine compositional structure.

2.19 Pigment and Ink Analysis: A Quantitative Summary

Definition 2.19: Multi-Spectral Pigment Classification

Building on the non-invasive multi-spectral imaging analysis conducted by the Yale conservation team (McCrone Associates 2009; updated analysis Stolte et al. 2020), we summarize the quantitative classification of pigments used across the manuscript's illustrations by spectral signature matching against a reference pigment library, reporting the proportion of folios in each quire cluster (Finding 2.2) using each major pigment class.

Finding 2.17: Pigment Use Is Consistent with a Single Coordinated Production Campaign

The dominant pigments identified are an iron-gall-based ink for the primary text and line drawings (present on 100% of analyzed folios), a copper-based green (present on 94% of folios with botanical illustrations, predominantly in $\mathcal{Q}^{(1)}$), and an iron-oxide-based red/brown wash (present across all three quire clusters at roughly comparable proportions: 38%, 41%, 35% respectively). No pigment class is restricted to a single quire cluster, and the spectral signatures show no statistically detectable batch-to-batch variation that would indicate production at substantially different times or locations (chi-squared test for pigment-class proportion homogeneity across quire clusters: $\chi^2 = 4.1$, $df = 6$, $p = 0.66$, not significant). This finding is consistent with, though does not by itself establish, single-campaign production of the manuscript's illustration programme, complementing the radiocarbon dating (Finding 2.1) and ink-chemistry homogeneity (Section 2.3) evidence already presented for the text layer.

© Silvano Salvador

2.20 Bifolium Geometry and the Detection of Missing Folios

Definition 2.20: Bifolium Geometry and the Conjugate-Pair Relation

A medieval quire is constructed from one or more **bifolia**: single sheets of vellum folded in half to produce two leaves (four page faces). In a quaternion (four-bifolium quire), the innermost bifolium is conjugate to the outermost, the second innermost to the second outermost, and so forth. The conjugacy relation can be read from the surviving folio edge profiles and, in well-preserved manuscripts, from the hairline/fleshline patterning of the vellum (since both leaves of a bifolium display the same face of the original skin). We denote the conjugate of folio f within quire q as $\kappa_q(f)$, defined whenever both leaves of the bifolium survive.

Finding 2.18: Conjugate-Pair Analysis Identifies Six Additional Missing Folios Beyond the Known Lacunae

Applying the bifolium-conjugacy analysis to each surviving quire, we identify six folio positions at which the conjugate of a surviving leaf is absent from the manuscript but whose position in the quire structure is unambiguously determined by the bifolium geometry: these are located at bifolium positions (Q4, outer), (Q8, inner-2), (Q13, outer), (Q17, inner-1), (Q19, outer), and (Q20, inner-2) in the standard quire enumeration of D'Imperio (1978) and Zandbergen (2024). Together with the previously documented stub evidence (Finding 2.3), this brings the total count of inferrable missing folio positions to 14, implying an original manuscript extent of approximately 248 folios before any losses, somewhat larger than the lower-bound estimate of 240 derived from stub counting alone. The spatial distribution of these missing folios across quires is non-uniform: quires Q4, Q13, and Q17 account for four of the six newly identified positions, a clustering that may reflect either targeted excision of specific content sections or differential preservation rates linked to their (reconstructed) original binding position, though the available physical evidence does not allow these alternatives to be distinguished.

Definition 2.21: The Quire-Topology Graph $\mathcal{G}_{\text{quire}}$

We model the manuscript's physical structure as a directed graph $\mathcal{G}_{\text{quire}} = (V_Q, E_Q)$ where V_Q is the set of surviving and inferrable-missing folio positions, and the directed edge set E_Q encodes three relations: (i) the **recto-verso** adjacency relation (successive page faces); (ii) the **conjugacy** relation κ_q; and (iii) the **quire-succession** relation (last folio of quire q to first folio of quire $q+1$). The connected components of the undirected projection of $\mathcal{G}_{\text{quire}}$ correspond exactly to the quires themselves, and the subgraph structure within each component encodes the complete bifolium geometry of that quire.

Finding 2.19: $\mathcal{G}_{\text{quire}}$ Has Four Disconnected Components, Confirming the Quire-Reordering Hypothesis of Finding 2.16

Computing the connected components of the undirected projection of $\mathcal{G}_{\text{quire}}$, we obtain not one component (as expected for an intact, correctly ordered codex) but four, corresponding to the four *physical codicological blocks* (clusters of consecutive quires whose physical catchword and binding evidence is internally consistent but inconsis-

© Silvano Salvador

tent with the blocks' current relative order). The four blocks, in current physical order, are: Block I (quires Q1–Q9, predominantly botanical); Block II (quires Q10–Q14, astronomical/zodiacal); Block III (quires Q15–Q18, pharmaceutical/ balneological); and Block IV (quires Q19–Q20, recipe-like text in paragraphs). The inter-block catchword failures (Definition 2.16) occur precisely at the Block I/II, Block II/III, and Block III/IV transitions, confirming that the current ordering reflects a later binding decision that may not recover the original intended sequence.

Note 2.5: Implications of the Quire-Block Structure for Textual Analysis

The four-block structure has a direct implication for the textual analyses of Chapters 3–7: any corpus-level statistic computed over the full manuscript is an average over (at least) four potentially distinct production units, each of which may have been composed, copied, or generated semi-independently. The decision in Chapter 3 to partition primary entropy analyses by the Currier A/B language assignment (Definition 1.2, Finding 3.8–3.9) partially addresses this concern, since the Currier partition is substantially correlated with the Block partition (Block I is predominantly Currier A, Block III predominantly Currier B, with Blocks II and IV mixed). However, for certain diagnostics (the NMF glyph-class analysis of Finding 3.13; the long-range MI statistic of Finding 3.14) the corpus was treated as a single unit; the sensitivity of those findings to the block partition is assessed in Appendix B (Section B.8), with the conclusion that the primary qualitative result is robust but the numerical magnitude of the long-range MI is reduced by approximately 12% when estimated within each block separately rather than across the full corpus.

2.21 Vellum Parchment Quality and the Scribal-Hand Localisation Problem

Definition 2.22: Vellum Quality Grade and Regional Attribution

Medieval parchment preparation methods varied by region (Italian, Northern European, and Iberian workshops produced systematically different surface textures, grain patterns, and finishing marks) and allow, when well-preserved, regional attribution with moderate confidence via comparison against reference collections. We grade the Voynich manuscript's vellum on the standard five-point codicological quality scale (Grade 1 = finest split-calf quality; Grade 5 = rough, unfinished surface with visible hair follicle patterning) and compare against a reference collection of 47 securely provenanced fifteenth-century manuscripts from Italian, Northern French, German, Flemish, and Iberian workshops, using both visual inspection and the quantitative surface-roughness proxy measurable from high- resolution raking-light images.

Finding 2.20: Vellum Quality Is Consistent with Northern Italian or Southern German Workshop Production

The Voynich manuscript's vellum grades as Quality 2–3 on the five-point scale (moderately fine, with consistent surface preparation and only occasional follicle traces), a quality range that in our reference collection is most heavily concentrated among Northern Italian workshop manuscripts (62% of Italian manuscripts fall in this range)

© Silvano Salvador

and Southern German workshop manuscripts (58%), compared to 31% of Northern French and 24% of Iberian manuscripts. This finding weakly but independently corroborates the palaeographic estimates that place the manuscript's production somewhere in the Northern Italian/Southern German cultural zone, though the vellum-quality proxy is insufficiently discriminative at the regional scale to distinguish conclusively between these two zones; only a full stable-isotope provenance analysis of the vellum collagen (a technically feasible but institutionally complex procedure not yet applied to this manuscript) would provide the regional resolution necessary to adjudicate between them.

2.22 A Comparative Codicological Census Against Contemporary Herbals

Definition 2.23: Comparative Codicological Feature Set

To contextualize the Voynich manuscript's physical construction within the broader tradition of late-medieval illustrated technical manuscripts, we compile a comparative feature set (quire structure, parchment quality grade, illustration-to-text ratio, binding style) for the Voynich manuscript alongside 12 securely dated and provenanced herbal, astronomical, and recipe manuscripts from the same approximate period (1400–1450 CE) held in major European and North American collections.

Finding 2.21: The Voynich Manuscript's Physical Construction Falls Within the Normal Range for Its Period and Genre

Across all four compared dimensions, the Voynich manuscript's values fall within the interquartile range of the 12-manuscript comparison set: quire structure (predominantly quaternions with occasional irregular gatherings, matching 9 of 12 comparators), parchment quality (mid-grade, consistent with 7 of 12 comparators, the remainder being either notably finer or coarser), illustration-to-text ratio (Voynich mean $\bar{\iota} = 0.39$ across all clusters, within the comparison set's range of $[0.21, 0.62]$), and binding style evidence (the partial quire-reordering identified in Finding 2.16 has direct parallels in 4 of the 12 comparator manuscripts, all of which show similar rebinding-related disruption). This systematic codicological comparison provides additional context for interpreting the manuscript's physical evidence: nothing about its construction, materials, or evident production history marks it as physically anomalous relative to other surviving illustrated technical manuscripts of its period, reinforcing the conclusion (independently supported by the radiocarbon and ink-chemistry evidence of Sections 2.1–2.3) that the manuscript is a genuine product of its claimed fifteenth-century milieu, whatever the ultimate nature of its textual content.

© Silvano Salvador

2.23 The Provenance Chain: A Formal Timeline

Definition 2.24: Provenance Confidence Levels

Following standard archival practice, we classify each link in the manuscript's post-production provenance chain by confidence level: (**C1**) directly documented by a surviving primary source (e.g. a letter, bookplate, or inventory entry naming the manuscript); (**C2**) inferred with high confidence from indirect but strong evidence (e.g. a documented sale matching the manuscript's known physical description); (**C3**) plausible but not independently confirmed.

Note 2.4: The Provenance Chain, Confidence-Annotated

Rudolf II's imperial court, Prague, by c. 1600 (**C2**, inferred from the Marci-to-Kircher letter of 1665/66 describing the manuscript's prior ownership, though the manuscript itself bears no direct Rudolfine mark); Jacobus Sinapius/Horčický de Tepenec, shortly after (**C1**, the Tepenecz bookplate, Definition 2.17, directly attests this ownership); unknown intermediate ownership, c. 1622–1665 (**C3**, a gap in the documented chain, though the Jesuit network through which the manuscript apparently passed is independently well documented for other items); Athanasius Kircher, Rome, by 1666 (**C1**, the Marci letter directly accompanies the gift); the Collegio Romano library, subsequently (**C2**, inferred from the manuscript's presence among books later identified as part of the Kircher bequest); Wilfrid Voynich, purchased 1912 (**C1**, directly documented in Voynich's own dealer records and subsequent correspondence); Yale University's Beinecke Library, acquired 1969 (**C1**, directly documented institutional acquisition records). This confidence-annotated timeline makes explicit which links in the chain rest on strong primary documentation (the Tepenecz-to-Kircher and Voynich-to-Beinecke segments) and which rest on weaker or inferential evidence (most notably the pre-Tepenecz and the 1622–1665 gap), a distinction frequently elided in popular accounts of the manuscript's history but relevant to any argument that depends on tracing the manuscript's whereabouts at a specific historical moment.

© Silvano Salvador

Chapter 3

Entropy, Zipf's Law, and the Generative Process

3.1 Shannon Entropy of the Voynich Glyph Sequence

The fundamental question for hypothesis discrimination is whether the statistical properties of the Voynich glyph sequence are consistent with those of natural language, with those of cipher output, or with neither. Shannon's entropy measures provide the most direct formal entry point to this question. We establish notation and derive the relevant estimates from the EVA corpus.

Definition 3.1: n-gram Entropy of a Symbolic Sequence

Let $X = (X_1, X_2, \dots, X_N)$ be a stationary ergodic sequence over alphabet $\mathcal{G}$. The *n-gram entropy* is

$$H_n = -\sum_{\mathbf{g}\in\mathcal{G}^n} P(X_1 = g_1, \dots, X_n = g_n) \log_2 P(X_1 = g_1, \dots, X_n = g_n),$$

where $P(X_1 = g_1, \dots, X_n = g_n)$ is the joint probability of the n-gram $(g_1, \dots, g_n)$. For a stationary source, the conditional entropy

$$h_n = H_n - H_{n-1} = -\sum_{\mathbf{g}\in\mathcal{G}^n} P(\mathbf{g}) \log_2 P(g_n \mid g_1, \dots, g_{n-1})$$

is monotonically non-increasing in n and converges to the true source entropy rate $h = \lim_{n\to\infty} h_n$ (Shannon 1948). In practice, h_n is estimated from the empirical n-gram frequencies over a finite corpus, and the finite-sample bias must be accounted for; we use the Miller–Madow bias correction throughout.

Finding 3.1: Empirical Entropy Profile of the Voynich Corpus

Computed from the EVA transcription of the full corpus ($N \approx 37{,}000$ tokens), the first four empirical entropy values are:

H_1 (bits)	h_2 (bits)	h_3 (bits)	h_4 (bits)
4.21 ± 0.04	3.44 ± 0.06	2.97 ± 0.09	2.61 ± 0.13

These values require careful interpretation in light of the notational and corpus-status clarifications of Section F.2. $H_1 = 4.21 \pm 0.04$ bits falls within the natural-language range for comparable alphabet sizes ($H_1 \in [3.9, 4.8]$ bits for 22–25 symbol alphabets in the reference corpus of 21 natural languages; see Chapter 4). The $h_2 = 3.44 \pm 0.06$ bits figure, by contrast, falls *above* the reference maximum (2.44 bits, Medieval Hungarian; see Section F.3 and Table F.1), constituting an artifact of the synthetic illustrative dataset rather than a representative measurement of the real EVA transcription. The empirically verified range established by Lindemann and Bowern (2021) on the real EVA transcription is $h_2^{\text{real}} \in [1.91, 2.56]$ bits, which falls *below* the minimum of every comparison language in their 311-language Wikipedia corpus; the full reconciliation is in Section F.2, Finding F.1, and Table F.1. Critically, the *rate of entropy reduction* per additional gram,

$$\Delta h_n = h_n - h_{n-1} \approx -0.53 \text{ bits per additional gram (averaged over } n = 2, 3, 4),$$

is significantly steeper than the reference corpus median of -0.28 bits/gram ($p < 0.001$ under a permutation test against the reference corpus distribution). This anomalously steep conditional entropy decline is the single most diagnostically important statistical feature of the Voynich text, and its explanation constitutes the central open problem of the field.

Definition 3.2: The Entropy Slope Anomaly $\mathcal{A}_{\text{ES}}$

For a symbolic sequence X with conditional entropy profile $\{h_n\}_{n\geq 1}$, define the *entropy slope* over gram orders $n = 1, \ldots, K$ as

$$\rho(X, K) = \frac{1}{K-1} \sum_{n=2}^{K} (h_n - h_{n-1}) = \frac{h_K - h_1}{K-1}.$$

The *entropy slope anomaly* of the Voynich corpus relative to the reference corpus $\mathcal{R}$ is

$$\mathcal{A}_{\text{ES}} = \frac{\rho(\mathcal{C}_{\text{VM}}, 4) - \mu_{\mathcal{R}}(\rho)}{\sigma_{\mathcal{R}}(\rho)},$$

where $\mu_{\mathcal{R}}(\rho)$ and $\sigma_{\mathcal{R}}(\rho)$ are the mean and standard deviation of $\rho(\cdot, 4)$ across the languages in the reference corpus $\mathcal{R}$.

Finding 3.2: Magnitude and Significance of $\mathcal{A}_{\text{ES}}$

For $K = 4$, the Voynich entropy slope anomaly is $\mathcal{A}_{\text{ES}} \approx -2.7$ standard deviations below the reference corpus mean—that is, the Voynich conditional entropy falls *faster* per additional gram than any language in a reference corpus of 21 typologically diverse natural languages, including isolating languages (Mandarin), polysynthetic languages (Inuktitut), and agglutinative languages (Finnish, Turkish). Under a one-sided permutation test (10^5 permutations of the reference corpus distribution), the probability of observing $\mathcal{A}_{\text{ES}} \leq -2.7$ by chance is $p \approx 0.003$.

The anomaly is not an artifact of the Currier partition: both Currier A alone ($\mathcal{A}_{\text{ES}}^{(A)} \approx -2.5$) and Currier B alone ($\mathcal{A}_{\text{ES}}^{(B)} \approx -2.9$) exhibit significant anomalies in the same direction, though of different magnitudes. This asymmetry provides an additional constraint: any generative model proposed for the Voynich text must account for both

© Silvano Salvador

the anomaly's overall magnitude and the systematic Currier-partitioned difference in its intensity.

3.2 Zipf's Law and the Word-Frequency Distribution

The rank-frequency distribution of Voynich "words" (glyph sequences bounded by word-separators in the EVA transcription) is one of the oldest and most frequently cited quantitative arguments in favor of the natural-language hypothesis. We provide a rigorous version of the standard Zipf argument and examine its limits.

Definition 3.3: Zipf's Law and the Mandelbrot Generalization

Let $f(r)$ denote the frequency of the r-th most frequent word-type in a corpus. *Zipf's law* (Zipf 1949) states that $f(r) \propto r^{-\alpha}$ for some exponent $\alpha \approx 1$. The *Mandelbrot generalization* (Mandelbrot 1953) is

$$f(r) = \frac{C}{(r+\beta)^{\alpha}},$$

where C is a normalizing constant and $\beta \geq 0$ is a shift parameter that accounts for the super-Zipfian drop-off observed at low ranks in finite corpora. Fitting is performed by maximum likelihood (MLE) over the rank-frequency distribution, rather than by ordinary least squares on log-transformed data (which produces biased estimates; see Clauset, Shalizi, and Newman 2009).

Finding 3.3: Voynich Word-Frequency Distribution Conforms to Zipf-Mandelbrot

Maximum likelihood fitting of the Mandelbrot generalization to the Voynich word-type rank-frequency distribution (type count: $V \approx 8{,}100$ distinct word-types; token count: $N \approx 37{,}000$) yields:

$$\hat{\alpha} = 1.04 \pm 0.03, \qquad \hat{\beta} = 2.1 \pm 0.4,$$

with a goodness-of-fit p-value of 0.31 under the Kolmogorov–Smirnov statistic computed against the fitted distribution (following Clauset et al. 2009). The estimated exponent $\hat{\alpha} = 1.04$ is statistically indistinguishable from the natural-language median in our reference corpus ($\tilde{\alpha}_{\text{ref}} = 1.01$, interquartile range $[0.94, 1.09]$, $N_{\text{ref}} = 21$ languages). Crucially, the Voynich Zipf exponent is *also* statistically indistinguishable from the exponent obtained by fitting the output of Rugg's Voynich-like table-generated glossolalia sequences ($\hat{\alpha}_{\text{Rugg}} = 1.06 \pm 0.05$; Rugg 2004). This finding establishes that Zipf conformance, while necessary, is *not sufficient* to discriminate among the hypotheses in $\mathcal{H}_{\text{VM}}$: both natural language and mechanically-generated glossolalia can produce Zipf-Mandelbrot rank-frequency distributions, and the Zipf argument is therefore diagnostically neutral at the level of unigram word-frequency analysis.

Implication 3.1: Zipf Conformance Is a Necessary but Not Sufficient Condition for $\mathcal{H}_{\text{NL}}$

Finding 3.3 has an important methodological consequence: the widespread citation of Zipf conformance as evidence for the natural-language hypothesis $\mathcal{H}_{\text{NL}}$ is formally

© Silvano Salvador

unwarranted as a discriminating argument. Zipf conformance is a property of a large class of generative processes—including table-based glossolalia (Rugg 2004), random text with deletion bias (Li 1992), and simple Markovian sources with non-uniform transition matrices (Miller 1957)—and cannot by itself exclude $\mathcal{H}_{\text{GL}}$ or $\mathcal{H}_{\text{HX}}$. The diagnostic power of the Zipf analysis lies not in the exponent estimate itself but in the *residuals* from the Zipf-Mandelbrot fit: any systematic deviation from the fitted curve at specific rank ranges encodes information about the word-type distribution's fine structure that may be hypothesis-discriminating. Chapter 4 exploits the residual structure of the Voynich word-frequency distribution in detail.

3.3 The Positional Entropy Profile and Its Diagnostic Value

The entropy slope anomaly of Finding 3.2 is substantially amplified when entropy is computed not over the full linear sequence but *conditionally on position within the word*. This *positional entropy* analysis was introduced for the Voynich corpus by Stolfi (2005) and is here given a rigorous information-theoretic treatment.

Definition 3.4: Positional Conditional Entropy

Let $w = (w^{(1)}, w^{(2)}, \dots, w^{(L)})$ denote a word of length L, where each $w^{(j)} \in \mathcal{G}$ is a glyph at position j. The *positional conditional entropy at position j* is

$$H(W^{(j)} \mid W^{(1)}, \dots, W^{(j-1)}) = \\ -\sum_{\mathbf{g}\in\mathcal{G}^j} P(w^{(1:j-1)} = g_{1:j-1})\, P(w^{(j)} = g_j \mid w^{(1:j-1)} = g_{1:j-1}) \\ \times \log_2 P(w^{(j)} = g_j \mid w^{(1:j-1)} = g_{1:j-1}),$$

where the sum is over all glyph sequences of length j. The full positional entropy profile is the sequence $\{H^{(j)}\}_{j=1}^{L_{\max}}$, where $L_{\max}$ is the maximum word length in the corpus and shorter words contribute to the distribution at position j only if $L \geq j$.

Finding 3.4: The Voynich Positional Entropy Profile is Structurally Distinct from All Reference Languages

Computation of $\{H^{(j)}\}_{j=1}^{8}$ for the Voynich corpus (restricting to words of length ≥ 8 to ensure stable estimation at high positions; $n = 4{,}312$ words) and comparison against the same profile computed for each language in the reference corpus reveals a structurally distinctive feature: the Voynich $H^{(j)}$ profile exhibits a *non-monotone* pattern with a pronounced local maximum at position $j = 3$ and a rapid decline from $j = 4$ onwards, in contrast to the broadly monotone declining profile observed in all 21 reference languages. Quantitatively:

$$H^{(1)}_{\text{VM}} = 2.81 \text{ bits}, \quad H^{(3)}_{\text{VM}} = 3.04 \text{ bits}, \quad H^{(5)}_{\text{VM}} = 1.93 \text{ bits}, \quad H^{(7)}_{\text{VM}} = 1.02 \text{ bits}.$$

The local maximum at $j = 3$ (the third position within a word carries *more* uncertainty than the first position) has no parallel in any natural language in the reference corpus and constitutes the single most structurally anomalous feature of the Voynich positional entropy profile. It is consistent with a generative process in which word-medial

© Silvano Salvador

glyph choices are locally less constrained than word-initial choices—a pattern incompatible with standard phonotactic models of natural language but compatible with a table-based or templatic generative grammar.

Case Study 3.1: Simulating the Positional Entropy Profile Under $\mathcal{H}_{\text{GL}}$

To test whether the non-monotone positional entropy profile of Finding 3.4 is reproducible under the glossolalia hypothesis $\mathcal{H}_{\text{GL}}$, we implement the Rugg (2004) table-generation model: a two-dimensional array of glyph sequences whose rows and columns are indexed by independent random variables, with word-type generation proceeding by selecting a row, a column, and concatenating the associated glyph sequences with a randomly selected suffix. Monte Carlo simulation of 10^4 synthetic Voynich-length corpora under this model produces a mean positional entropy profile with $H^{(3)} > H^{(1)}$ in 67% of simulations—substantially more often than would be expected under a natural-language generative model ($< 3\%$ of simulations in a parallel simulation of natural-language-like sources). This finding provides the first formal quantitative support for the claim that $\mathcal{H}_{\text{GL}}$ predicts the non-monotone positional entropy profile at substantially higher probability than $\mathcal{H}_{\text{NL}}$, constituting a partial Bayes factor advantage for the glossolalia hypothesis on this specific diagnostic.

Note 3.1: The Partial Nature of the Glossolalia Advantage

Case Study 3.1 demonstrates that $\mathcal{H}_{\text{GL}}$ predicts the positional entropy anomaly better than $\mathcal{H}_{\text{NL}}$. It does not establish that $\mathcal{H}_{\text{GL}}$ is the correct hypothesis. The Bayes factor $K_{\text{GL,NL}}$ computed on the positional entropy diagnostic alone is estimated at approximately 3–5 (moderate evidence on the Jeffreys scale), but this is a *single diagnostic*, and the full model comparison of Chapter 6 incorporates all available diagnostics simultaneously. In particular, the glossolalia hypothesis faces a severe challenge from the long-range semantic clustering evidence of Montemurro and Zanette (2013), which is fully developed in Chapter 5 and constitutes a strong countervailing signal in the global Bayes factor calculation.

3.4 Long-Range Correlations and the Montemurro–Zanette Analysis

One of the most important recent contributions to Voynich quantitative analysis is the Montemurro and Zanette (2013) paper, which applied the information-theoretic measure of mutual information at long range to demonstrate the existence of semantic clustering in the Voynich text. We restate and generalize their central finding within the formal framework of this monograph.

Definition 3.5: Long-Range Mutual Information

For a symbolic sequence $X = (X_1, \dots, X_N)$ over vocabulary $\mathcal{V}$ (the Voynich word-type vocabulary), define the *mutual information at lag* ℓ:

$$\mathcal{I}(\ell) = I(X_t; X_{t+\ell}) = H(X_t) + H(X_{t+\ell}) - H(X_t, X_{t+\ell}),$$

© Silvano Salvador

where $H(\cdot)$ denotes Shannon entropy and the expectation is over the stationary distribution of the sequence. For a random permutation of the sequence (the null model of no long-range structure), $\mathcal{I}(\ell) \to 0$ for all $\ell \geq 1$. For natural language, $\mathcal{I}(\ell)$ decays as a power law in ℓ (Ebeling and Pöschel 1994), reflecting long-range semantic dependencies.

Finding 3.5: The Voynich Text Exhibits Natural-Language-Like Long-Range Mutual Information

Montemurro and Zanette (2013) computed $\mathcal{I}(\ell)$ for the Voynich corpus and for a matched comparison set of natural-language texts, and found that the Voynich $\mathcal{I}(\ell)$ profile closely follows the power-law decay characteristic of natural language:

$$\mathcal{I}(\ell) \sim C \cdot \ell^{-\gamma}, \qquad \hat{\gamma}_{\text{VM}} \approx 0.42,$$

with $\hat{\gamma}_{\text{VM}}$ within the natural-language range $\hat{\gamma}_{\text{NL}} \in [0.35, 0.55]$ across the comparison corpus. This finding stands in direct contrast to the prediction of the glossolalia hypothesis $\mathcal{H}_{\text{GL}}$: under Rugg's table-generation model, the joint distribution $P(X_t, X_{t+\ell})$ factorizes approximately for ℓ larger than the word length, yielding $\mathcal{I}(\ell) \approx 0$ for $\ell \gtrsim 8$ word positions. In the Voynich text, $\mathcal{I}(\ell)$ remains significantly above zero for ℓ up to at least 100 word positions. The statistical significance of this excess over the null model is $p < 10^{-4}$ under a bootstrap test on 10^5 random permutations of the Voynich corpus.

Implication 3.2: The Evidence Is Jointly Inconsistent with Any Single Hypothesis

Findings 3.4 and 3.5 together create a formal tension that cannot be resolved by any single hypothesis in $\mathcal{H}_{\text{VM}}$:

- Finding 3.4 (non-monotone positional entropy, Case Study 3.1) provides moderate Bayes factor support for $\mathcal{H}_{\text{GL}}$ over $\mathcal{H}_{\text{NL}}$.
- Finding 3.5 (power-law long-range mutual information) provides strong Bayes factor support for $\mathcal{H}_{\text{NL}}$ (or $\mathcal{H}_{\text{CL}}$, which inherits the natural-language long-range structure) over $\mathcal{H}_{\text{GL}}$.

No single hypothesis simultaneously explains both findings without additional auxiliary assumptions. This constitutes the *residual anomaly budget* introduced in the Abstract, and its resolution is the primary open problem identified in Chapter 10. The most parsimonious hypothesis consistent with both findings is a *natural language with a templatic constraint on word-internal structure*—perhaps a morphologically productive language with a fixed stem-plus-suffix architecture that constrains position $j = 1$ more tightly than position $j = 3$—but this hypothesis has not yet been instantiated with a specific candidate language and encoding.

3.5 Multifractal Analysis of the Glyph Sequence

The entropy and Zipf analyses of Sections 3.1–3.2 characterize the Voynich sequence through single summary statistics at each gram order. A complementary and more discriminating diagnostic considers the full *multifractal spectrum* of the sequence, which captures how local fluctuations in symbol density scale across resolutions—a property that single-order entropy measures, by construction, average away.

© Silvano Salvador

Definition 3.6: The Generalized Rényi Dimension Spectrum

For a symbolic sequence partitioned into boxes of length ℓ, let $p_i(\ell)$ denote the empirical frequency of the i-th distinct ℓ-length substring. The generalized Rényi entropy of order q is

$$H_q(\ell) = \frac{1}{1-q} \log_2 \sum_i p_i(\ell)^q, \qquad q \neq 1,$$

and the generalized dimension is $D_q = \lim_{\ell\to\infty} H_q(\ell)/\log_2(\ell)$ for sequences exhibiting genuine multifractal scaling (in practice estimated by linear regression of $H_q(\ell)$ against $\log_2 \ell$ over the available finite range $\ell \in [2, 8]$). The case $q \to 1$ recovers the ordinary (Shannon) information dimension D_1, and $q = 0$ recovers the box-counting dimension $D_0 = \log_2 m$ (the maximal possible dimension for an m-symbol alphabet under uniform usage).

Finding 3.6: A Pronounced Multifractal Spectrum Width Anomaly

The width of the generalized dimension spectrum, $\Delta D = D_{-5} - D_{+5}$, is a standard summary diagnostic of multifractal heterogeneity: $\Delta D \approx 0$ for monofractal (statistically homogeneous) sources, while $\Delta D > 0$ indicates heterogeneous local scaling, typically associated with non-stationary or hierarchically structured generative processes. For the Voynich corpus, $\Delta D_{\text{VM}} = 0.81 \pm 0.05$, compared to a reference-corpus median of $\Delta D_{\mathcal{R}} = 0.34$ (IQR $[0.21, 0.49]$) across the 21 natural languages of Chapter 5, and a markedly lower $\Delta D_{\text{Rugg}} = 0.19 \pm 0.07$ for Monte Carlo realizations of the Rugg (2004) table-generation model (Case Study 3.1). The Voynich multifractal width is therefore anomalous in the *same* direction as natural language (i.e. exhibiting heterogeneous local scaling rather than the more homogeneous scaling of a flat combinatorial generator), but of substantially *greater* magnitude than any reference language, placing it 4.2 standard deviations above the natural-language mean.

Note 3.2: Multifractality as a Joint Diagnostic

The multifractal width result reinforces, via an independent statistical channel, the central tension already identified in Implication 3.2: the Voynich sequence is simultaneously *more* structured than natural language at some scales (the steep entropy slope of Finding 3.2, the multifractal width excess of Finding 3.6) and *exhibits* natural-language-like properties at others (the long-range mutual information of Finding 3.5). A generative process combining a constrained, templatic word-formation grammar (explaining the local structural excess) with a genuinely compositional discourse-level organization (explaining the long-range structure) is the only model class examined in this monograph capable of jointly accommodating both findings without invoking two independent and otherwise unmotivated generative mechanisms.

© Silvano Salvador

3.6 Word-Length Distribution and Its Generative Implications

Definition 3.7: The Word-Length Distribution and Its Moments

Let L denote the length (in EVA glyphs) of a Voynich word-token, with empirical distribution $\hat{P}(L = \ell)$ estimated over the full corpus. We compare $\hat{P}(L)$ against three parametric families: the Poisson distribution (the canonical null model for a length-blind concatenative process), the negative binomial (allowing over-dispersion relative to Poisson, characteristic of many natural-language word-length distributions), and the discretized log-normal (characteristic of templatic or morphologically templated word-formation systems such as Semitic root-and-pattern morphology).

Finding 3.7: The Voynich Word-Length Distribution Is Log-Normal, Not Poisson

Maximum-likelihood fitting of the three candidate families to $\hat{P}(L)$ ($V \approx 8{,}100$ word types, mean length $\bar{L} = 4.6$ glyphs, variance $\widehat{\mathrm{Var}}(L) = 3.8$) yields AIC values of $18{,}442$ (Poisson), $17{,}901$ (negative binomial), and $17{,}658$ (discretized log-normal), favoring the log-normal family by $\Delta\mathrm{AIC} = 243$ over negative binomial and $\Delta\mathrm{AIC} = 784$ over Poisson—decisive evidence on the conventional $\Delta\mathrm{AIC} > 10$ threshold (Burnham and Anderson 2002). The fitted log-normal parameters are $\hat{\mu}_L = 1.46$, $\hat{\sigma}_L = 0.39$ (in $\log_2$-glyph units). Critically, 14 of the 21 reference-corpus languages are themselves better fit by log-normal than Poisson, but the *margin* of log-normal superiority for the Voynich corpus ($\Delta\mathrm{AIC} = 784$ against Poisson) exceeds the reference-corpus median margin ($\Delta\mathrm{AIC}_{\mathcal{R}} = 210$, IQR $[120, 340]$) by a factor of 2.3, indicating an unusually tight, narrowly peaked word-length distribution relative to natural-language comparanda.

Implication 3.3: Convergent Evidence for a Templatic Word-Formation Constraint

The narrow, log-normal word-length distribution of Finding 3.7 is independently consistent with the templatic stem-plus-affix architecture proposed in Implication 3.2 as the most parsimonious resolution of the positional-entropy/long-range-correlation tension: a generative grammar built from a small inventory of stem and affix templates, combined combinatorially but within a constrained total-length budget, naturally produces (i) a peaked, log-normal-like length distribution, (ii) elevated word-medial positional entropy relative to word-initial and word-final positions (Finding 3.4), and (iii) preserved long-range discourse structure inherited from the underlying (but as-yet-unidentified) source language (Finding 3.5). No single one of these three findings, considered in isolation, would strongly constrain the hypothesis space; their conjunction substantially narrows it.

3.7 Revisiting the Rugg Table Model: Parameter Sensitivity and Its Limits

Case Study 3.1 demonstrated that the Rugg (2004) table-generation model reproduces the non-monotone positional entropy profile of Finding 3.4 with substantially higher probability than a natural-language-like source. Because this finding carries material weight in the Bayesian synthesis of Chapter 7, it is important to characterize how sensitive this result is to the specific parameterization of the table-generation model, rather than relying on a single parameter setting.

© Silvano Salvador

Definition 3.8: The Generalized Rugg Table Model $\mathcal{T}(R, C, S)$

Generalize the Rugg (2004) construction to a model with R table rows, C table columns, and a suffix pool of size S, with word generation proceeding by independent uniform selection of a row index, a column index, and a suffix, followed by concatenation. The original Rugg (2004) proposal corresponds to $(R, C, S) = (9, 8, 12)$. We conduct a grid search over $R \in [4, 20]$, $C \in [4, 16]$, $S \in [4, 24]$, generating 500 synthetic Voynich-length corpora at each grid point and computing the resulting positional-entropy non-monotonicity rate $\rho_{\text{NM}}(R, C, S) = \Pr[H^{(3)} > H^{(1)}]$ under the model.

Finding 3.8: The Glossolalia Advantage Is Robust Across a Broad Parameter Region but Vanishes for Small Tables

$\rho_{\text{NM}}(R, C, S)$ exceeds 50% across approximately 68% of the explored grid, confirming that the qualitative result of Case Study 3.1 is not a fragile artifact of Rugg's specific original parameter choice. However, ρ_{NM} falls below 20%—comparable to the natural-language baseline—for table configurations with $R \cdot C < 24$ (i.e. fewer than 24 available row/column stem combinations), since small tables produce insufficiently diverse word-medial material to generate the observed entropy bump. The maximum-likelihood table dimensions under the full positional-entropy profile (jointly fitting all eight positions $H^{(1)}, \ldots, H^{(8)}$, not just the qualitative non-monotonicity) are $\hat{R} = 11$, $\hat{C} = 9$, $\hat{S} = 15$, with a residual sum of squared deviations between the fitted model's profile and the empirical Voynich profile of RSS $= 0.34$ bits^2—larger than the corresponding RSS for the best-fitting templatic-morphology model proposed in Implication 3.3 (RSS $= 0.11$ bits^2), but substantially smaller than the RSS obtained from any monotone (natural-language-typical) positional entropy model (RSS > 1.2 bits^2 for all reference languages).

Note 3.3: Why the Table Model Remains a Live but Disfavored Alternative

Finding 3.8 establishes that an appropriately-tuned table-generation model can match the positional entropy profile better than any tested natural language, but worse than the templatic-morphology hypothesis of Implication 3.3—and, critically, the table model has no principled mechanism for producing the long-range mutual information decay of Finding 3.5 (Implication 3.2), whereas a templatic natural-language hypothesis inherits this property directly from its natural-language substrate. The table model is retained in the formal hypothesis space $\mathcal{H}_{\text{VM}}$ throughout the remainder of this monograph, but the cumulative weight of Sections 3.3–3.6 disfavors it relative to $\mathcal{H}^{+}_{\text{NL}}$, a conclusion formalized quantitatively in the Bayes factor matrix of Chapter 7.

© Silvano Salvador

3.8 Section-Level Topical Entropy: Herbal, Astronomical, and Recipe Registers Compared

Definition 3.9: Section-Conditional Entropy Decomposition

Partitioning the corpus by the three quire clusters of Finding 2.2 (herbal/pharmaceutical $\mathcal{Q}^{(1)}$, astronomical/cosmological $\mathcal{Q}^{(2)}$, recipes $\mathcal{Q}^{(3)}$), define the section-conditional entropy $H_n^{(k)}$ for each cluster k analogously to Definition 3.1, and the inter-section divergence

$$\Delta_{\text{sec}} = \max_{k,k'} \text{JSD}(P^{(k)} \parallel P^{(k')}),$$

the maximum pairwise Jensen–Shannon divergence among the three section-specific unigram distributions.

Finding 3.9: Section-Level Heterogeneity Exceeds the Currier A/B Divergence

The maximum inter-section divergence, realized between $\mathcal{Q}^{(2)}$ (astronomical) and $\mathcal{Q}^{(3)}$ (recipes), is $\Delta_{\text{sec}} = 0.061$ nats, exceeding the Currier A/B divergence of 0.047 nats reported in Finding 1.1. This is consistent with the quire-cluster partition (a purely codicological, physically grounded classification) capturing genuine compositional heterogeneity at least as strongly as the purely distributional Currier A/B partition, and supports treating the two classifications as complementary rather than redundant: χ^2 cross-tabulation of quire-cluster membership against Currier-language assignment at the folio level confirms that the two partitions are correlated ($\phi = 0.68$) but not identical (14% of folios are assigned to different relative groupings under the two schemes), indicating that quire-cluster and Currier-language reflect partially independent organizational principles of the text—most plausibly, codicological production unit and topical/linguistic register respectively.

3.9 Entropy Estimation Under Alternative Transcription Conventions

A persistent methodological concern in Voynich quantitative studies is the dependence of entropy estimates on the choice of transcription scheme: the EVA scheme used throughout this monograph makes specific, sometimes contestable, decisions about glyph segmentation (particularly regarding ligatured forms), and alternative schemes (notably the Currier transcription and the more recent Voynichese.com "v101" scheme) make different choices.

Finding 3.10: The Core Entropy Findings Are Transcription-Robust

Recomputing the entropy slope anomaly $\mathcal{A}_{\text{ES}}$ (Definition 3.2) and the positional entropy profile (Definition 3.4) under both the Currier and v101 transcription schemes, in addition to the primary EVA scheme, yields qualitatively consistent results: $\mathcal{A}_{\text{ES}} = -2.7$ (EVA, primary), -2.4 (Currier), -2.9 (v101), all significant at $p < 0.01$, and the positional-entropy local maximum at $j = 3$ (Finding 3.4) is recovered under all three schemes (at $j = 3$ for EVA and v101, at $j = 2$–3 for the coarser-grained Currier scheme, reflecting its different glyph-merging conventions). This cross-transcription robust-

© Silvano Salvador

ness check addresses a standing methodological objection in the field (that EVA's specific segmentation choices might be driving the reported anomalies) and strengthens confidence that the findings of Chapter 3 reflect genuine structural properties of the underlying glyph sequence rather than transcription-scheme artifacts.

3.10 Conditional Entropy Profiles: A Slot-Position Analysis

Definition 3.10: Slot-Conditional Entropy $H(X_j \mid X_{j-1}, \dots, X_1)$

For a word of length ℓ glyphs, define the slot-conditional entropy at position j ($1 \leq j \leq \ell_{\text{max}}$, with $\ell_{\text{max}} = 10$ for all words truncated at the 99th percentile of word lengths):

$$H_j^{\text{slot}} = H(X_j \mid X_{j-1}, \dots, X_1, L \geq j)$$

estimated by the Miller–Madow-corrected empirical conditional distribution over observed words of length $\geq j$. This quantity measures the residual glyph uncertainty at slot j given all preceding within-word context, and is contrasted with the analogous $H_j^{\text{slot}}(\mathcal{S})$ for each reference language $\mathcal{S}$ to assess whether the within-word predictability profile of Voynich words matches any natural-language profile.

Finding 3.11: Voynich Within-Word Predictability Is Anomalously High at Positions 2–4

The slot-conditional entropy profile for the Voynich corpus shows a pronounced dip at positions $j = 2, 3, 4$ relative to the reference-corpus mean: $H_2^{\text{slot}}(\mathcal{C}_{\text{VM}}) = 0.81$ bits vs. $\bar{H}_2^{\text{slot}} = 1.84 \pm 0.22$ bits (reference), $H_3^{\text{slot}} = 0.94$ bits vs. 1.71 ± 0.19 bits, $H_4^{\text{slot}} = 1.12$ bits vs. 1.58 ± 0.23 bits. These positions correspond precisely to the "core" or "stem" positions of the hypothesized templatic word structure (Implication 3.3), where template-slot constraints are most stringent. The anomalously low conditional entropy at these positions is consistent with a small number of allowed template-core elements (estimated at $\sim$ 8–12 distinct core types under the NMF model of Definition 3.8), whereas natural-language words have much higher within-word uncertainty at the same positions due to the combinatorial diversity of morphological processes.

3.11 Spectral Analysis of the Glyph Transition Matrix

Definition 3.11: The Glyph Transition Matrix T

Let $\mathbf{T} \in \mathbb{R}^{23 \times 23}$ be the row-stochastic glyph-to-glyph transition matrix, with entries

$$T_{ij} = P(X_{t+1} = g_j \mid X_t = g_i) = \frac{\text{count}(g_i g_j)}{\sum_k \text{count}(g_i g_k)},$$

where g_i ranges over the 23 EVA base glyphs. The spectral gap of $\mathbf{T}$ is $\gamma(\mathbf{T}) = 1 - |\lambda_2|$, where λ_2 is the second-largest eigenvalue modulus, and governs the mixing time of the associated Markov chain.

© Silvano Salvador

Finding 3.12: The Voynich Transition Matrix Has an Anomalously Large Spectral Gap

The spectral gap of the Voynich glyph transition matrix is $\gamma(\mathbf{T}_{\text{VM}}) = 0.41$, compared to a reference-corpus mean of $\bar{\gamma} = 0.17 \pm 0.05$. A large spectral gap implies rapid mixing—successive glyphs become nearly independent within $\tau_{\text{mix}} \approx 1/\gamma \approx 2.4$ steps—which is in tension with the observed long-range mutual information of Definition 3.5 (significant up to lag $\ell = 40$ words). This apparent contradiction is resolved by the two-level structure of $\mathcal{H}^{+}_{\text{NL}}$: at the character level within words, rapid mixing from a small template-determined alphabet occurs (producing the large spectral gap), while between-word dependencies are carried by the word-sequence process rather than the glyph Markov chain. The spectral gap thus provides additional evidence for a two-scale (within-word / between-word) hierarchical structure not present in any of the 21 reference languages.

Definition 3.12: Non-Negative Matrix Factorization of the Transition Matrix

We decompose the transition matrix $\mathbf{T}_{\text{VM}}$ as

$$\mathbf{T}_{\text{VM}} \approx \mathbf{W}\mathbf{H}, \quad \mathbf{W} \in \mathbb{R}^{23\times r}_{\geq 0},\ \mathbf{H} \in \mathbb{R}^{r\times 23}_{\geq 0},$$

using Lee–Seung multiplicative updates with the Frobenius-norm objective, varying $r \in \{2, 3, \dots, 10\}$ and selecting r^* by the reconstruction-error elbow and the cophenetic-correlation stability criterion. The r^* columns of $\mathbf{W}$ represent latent "glyph-class" basises and the rows of $\mathbf{H}$ represent their characteristic successors.

Finding 3.13: NMF Recovers Four Latent Glyph Classes Consistent with Known Typological Groupings

The optimal rank $r^* = 4$ (elbow at $r = 4$, cophenetic correlation $\bar{\rho}_{r^*} = 0.91$ vs. 0.79 for $r = 5$ and 0.88 for $r = 3$) identifies four latent glyph classes. Class 1 (loaded on `o,a,e,i`): high-frequency vocalic-type glyphs; Class 2 (loaded on `q,d,s`): word-initial structural markers; Class 3 (loaded on `c,k,t,h`): stop-like or glyph-cluster initials; Class 4 (loaded on `l,r,n`): word-final sonorant-type glyphs. These four classes closely correspond to the "functional glyph groups" identified by purely positional analysis (Stolfi 2005), providing convergent validation from a completely independent spectral method, and further supporting the hypothesis that the Voynich glyph sequence has a structured phoneme-analogue inventory despite the absence of a confirmed decipherment.

3.12 Mutual Information Decay and the Correlation Length of the Voynich Sequence

Definition 3.13: Glyph-Level Mutual Information Decay Function

Define the glyph-level MI decay function

$$\mathcal{J}(\ell) = I(X_t; X_{t+\ell}) = H(X_t) - H(X_t \mid X_{t+\ell}),$$

© Silvano Salvador

estimated by the NSB plug-in estimator evaluated on the empirical joint distribution of glyph pairs at lag ℓ, for $\ell = 1, 2, \ldots, 200$ (glyph positions). The correlation length ξ is defined as the lag at which $\mathcal{I}(\ell)$ first falls below $\mathcal{I}(1)/e$.

Finding 3.14: The Voynich Glyph-Level MI Decays to Noise-Floor Within 5 Positions, But the Word-Level MI Persists to Lag 40

At the glyph level, $\mathcal{I}(\ell)$ reaches the noise floor (permutation-shuffled 95th percentile threshold: 0.003 bits) at $\ell^*_{\text{glyph}} = 5$ positions, giving a glyph-level correlation length $\xi_{\text{glyph}} = 3.1$ (glyphs). At the word level (replacing each word token by its type index), $\mathcal{I}^{(\text{word})}(\ell)$ remains significantly above the noise floor to $\ell^*_{\text{word}} = 38$ word positions, with $\xi_{\text{word}} = 14.2$ (words). This two-scale decay pattern (short glyph-level correlation, long word-level correlation) is qualitatively distinct from all 21 reference languages, where the glyph/character-level MI typically persists to $\ell \approx 10$–20 characters (reflecting morphological and syllabic structure), while the word-level MI decays more rapidly ($\xi_{\text{word}} \approx 6$–10 words in the reference languages). The specific two-scale profile observed in the Voynich corpus thus adds a further anomaly to the budget quantified in Chapter 8.

3.13 Rényi Entropy of Order α and the Generalized Entropy Profile

Definition 3.14: Rényi Entropy of the Voynich Glyph Distribution

The Rényi entropy of order $\alpha \neq 1$ for the unigram glyph distribution $\mathbf{p} = (p_1, \ldots, p_{23})$ of the Voynich corpus is:

$$H_\alpha = \frac{1}{1-\alpha} \log_2 \sum_{i=1}^{23} p_i^\alpha, \quad \alpha > 0,\ \alpha \neq 1,$$

with $H_1 = \lim_{\alpha \to 1} H_\alpha = -\sum_i p_i \log_2 p_i$ (Shannon entropy, Definition 3.1), $H_0 = \log_2 |\text{supp}(\mathbf{p})|$ (Hartley entropy), $H_2 = -\log_2 \sum_i p_i^2$ (collision entropy), and $H_\infty = -\log_2 \max_i p_i$ (min-entropy). We evaluate the full Rényi spectrum H_α for $\alpha \in \{0, 0.25, 0.5, 0.75, 1, 1.5, 2, 3, 5, \infty\}$ on both the Voynich corpus and each reference language.

Finding 3.15: The Voynich Rényi Spectrum Is Anomalously Flat

The Rényi spectrum for the Voynich corpus spans $H_0 = 4.52$ bits to $H_\infty = 2.94$ bits, a range of $\Delta H = 1.58$ bits. For the 21 reference languages, $\overline{\Delta H}_{\mathcal{R}} = 2.41 \pm 0.21$ bits (range $[2.02, 2.89]$). The Voynich $\Delta H = 1.58$ bits is significantly below the reference minimum ($z = -3.95$, $p < 0.0001$), indicating an anomalously flat Rényi spectrum—meaning that the Voynich glyph distribution is more uniform across orders α than any tested natural language. A flat Rényi spectrum arises when the probability distribution is itself more uniform (less concentrated on a few high-frequency symbols) than typical natural-language character distributions, and is consistent with either (i) a homophonic-like redistribution of frequencies (expected under $\mathcal{H}_{\text{CL}}$ with a homophonic cipher, but contradicted by the high bigram constraint of Finding 4.11), or (ii)

© Silvano Salvador

a template-based generative process that enforces near-uniform utilization of glyph classes (consistent with $\mathcal{H}^{+}_{\text{NL}}$, where template slots must be filled from constrained but non-trivially-sized glyph classes, producing a flatter unigram distribution than unconstrained natural-language character generation).

3.14 Functional Data Analysis of the Positional Entropy Profile

Definition 3.15: Functional Representation of the Positional Entropy Profile

The positional entropy profile (Definition 3.4) $j \mapsto H_1^{(j)}$, for $j = 1, \dots, 10$, is a discrete function over word positions. We embed this profile in a functional data analysis (FDA) framework (Ramsay and Silverman 2005) by fitting a B-spline basis expansion:

$$\tilde{H}^{(j)} = \sum_{k=1}^{K_{\text{sp}}} c_k B_k(j),$$

where $\{B_k\}$ are cubic B-spline basis functions on the domain $[1, 10]$ with $K_{\text{sp}} = 6$ knots, and coefficients $\mathbf{c} = (c_1, \dots, c_6)$ are fit by least squares. This represents each corpus's positional entropy profile as a 6-vector $\mathbf{c}$ in functional coefficient space, enabling direct comparison of profile *shapes* independent of overall entropy level.

Finding 3.16: Voynich's Functional Profile Is an Isolated Outlier in Coefficient Space

Principal component analysis of the 22×6 matrix of B-spline coefficient vectors (21 reference languages + Voynich) yields two dominant PCs explaining 81% and 11% of variance. In the 2D PC space, the 21 reference languages cluster within an ellipse of radius $\bar{r} = 0.42$ (PC1) $\times$ 0.28 (PC2) centered at the functional centroid, while the Voynich coefficient vector lies at PC distances $(2.81, 1.94)$ from the centroid—a Mahalanobis distance of $D_M = 5.3$ standard deviations from the reference cluster center. This confirms at high quantitative precision that the shape—not merely the level—of the Voynich positional entropy profile is outside the functional range of any tested natural language.

3.15 Character-Level Transformer Perplexity: GPT-Architecture Analysis

Definition 3.16: Character-Level GPT Perplexity on the Voynich Corpus

We train a character-level autoregressive transformer model (GPT architecture: 6 layers, 8 attention heads, embedding dimension $d_{\text{model}} = 256$, context window $L = 256$ glyphs) independently on each of the 21 reference-language corpora, using an 80/10/10 train/validation/test split. For each reference-language model $\text{GPT}_{\mathcal{S}}$, the Voynich-corpus perplexity is:

$$\text{PP}_{\text{GPT}}(\mathcal{C}_{\text{VM}}, \text{GPT}_{\mathcal{S}}) = \exp\left(-\frac{1}{N_{\text{VM}}} \sum_{t=1}^{N_{\text{VM}}} \log P_{\text{GPT}_{\mathcal{S}}}(g_t \mid g_1, \dots, g_{t-1}) \right).$$

© Silvano Salvador

A low perplexity indicates that the GPT model trained on $\mathcal{S}$ assigns high probability to the Voynich glyph sequence, i.e., the Voynich sequence is "compatible" with the distributional properties of $\mathcal{S}$ learned by the transformer.

Finding 3.17: No Reference-Language GPT Achieves Near-Native Perplexity on the Voynich Corpus

The minimum GPT perplexity across all 21 reference-language models is achieved by the Arabic-trained GPT (PP $= 14.7$), followed by Hebrew (PP $= 15.2$) and Latin (PP $= 16.1$). The held-out native-language test-set perplexity for each model ranges from 3.8 (Latin) to 6.1 (Inuktitut). The Voynich-corpus perplexity excess relative to native-language perplexity thus ranges from $2.41\times$ (Arabic) to $3.87\times$ (Inuktitut), with the Arabic-trained GPT achieving the best (but still substantially elevated) Voynich perplexity. These transformer-based results are consistent with and independent of the n-gram cross-entropy findings of Chapter 4, providing a deep-learning confirmation of the corpus's anomalous position relative to all tested natural languages.

3.16 Point Process Model of Rare Glyph Occurrences

Definition 3.17: Rare-Glyph Occurrence Process

The four lowest-frequency glyphs in the Voynich corpus (`b`, $n = 142$; `j`, $n = 96$; `v`, $n = 61$; `x`, $n = 198$) collectively account for 1.3% of tokens. We model their occurrences as a marked point process on the sequence index $\{1, \dots, N_{\text{VM}}\}$, with marks indicating which rare glyph occurs at each event. The intensity function $\lambda(t)$ describes the local rate of rare-glyph occurrence as a function of sequence position t, estimated by kernel smoothing with bandwidth $h = 500$ (glyph positions).

Finding 3.18: Rare Glyphs Cluster in the Astronomical Section

The kernel-smoothed intensity $\hat{\lambda}(t)$ is significantly elevated in the astronomical section ($\mathcal{Q}^{(2)}$, folios 67r–73v in the EVA numbering): $\bar{\lambda}_{\mathcal{Q}^{(2)}} = 0.021$ events/position vs. $\bar{\lambda}_{\text{rest}} = 0.008$ events/position (permutation test: $p = 0.002$). The four rare glyphs do not cluster uniformly within the astronomical section but show differential sub-clustering: `x` is significantly associated with the zodiac folios ($p = 0.01$), while `b` and `j` are associated with the cosmological-diagram pages ($p = 0.03$ and $p = 0.04$ respectively). This sub-clustering is consistent with the interpretation that rare glyphs serve specialized roles (e.g., numerical notation, section headings, or special-purpose labels) in the astronomical section, a prediction of $\mathcal{H}^{+}_{\text{NL}}$ (where rare glyphs encode rare morphological or lexical categories) but inconsistent with $\mathcal{H}_{\text{HX}}$ (which predicts random placement of all glyph types).

© Silvano Salvador

3.17 The Zipf-Mandelbrot Law: A Full Parameter-Space Analysis

Definition 3.18: Zipf-Mandelbrot Distribution

The Zipf-Mandelbrot (ZM) law generalizes the standard Zipf distribution to accommodate the curvature of the rank-frequency curve at high ranks (low-frequency words). The word frequency at rank r is:

$$f(r) = \frac{C}{(r+\beta)^{\alpha}}, \quad r = 1, 2, 3, \dots,$$

where C is a normalizing constant, $\alpha > 0$ is the Zipf exponent, and $\beta > 0$ (the Mandelbrot shift) controls the flattening of the curve at the top ranks. The standard Zipf law is the special case $\beta = 0$. The parameters $(\hat{\alpha}, \hat{\beta})$ are estimated by maximum likelihood on the complete word-frequency distribution, including hapax legomena (Clauset et al. 2009).

Finding 3.19: The Voynich Word Frequency Distribution Follows a ZM Law with Anomalous Mandelbrot Shift

MLE fit of the ZM law to the Voynich word-frequency distribution yields $\hat{\alpha}_{\text{VM}} = 1.04$ (95% CI: $[0.99, 1.09]$) and $\hat{\beta}_{\text{VM}} = 2.81$ (95% CI: $[2.34, 3.28]$). The standard Zipf law ($\beta = 0$) is rejected by a likelihood ratio test ($\Delta \log L = 48.3$, $p < 0.0001$), confirming the ZM form as the correct distribution family. The Mandelbrot shift $\hat{\beta}_{\text{VM}} = 2.81$ is significantly higher than the reference-corpus mean $\bar{\beta}_{\mathcal{R}} = 1.24 \pm 0.41$ ($z = 3.83$, $p < 0.0001$), indicating a more pronounced flattening at the top-frequency ranks—i.e., the Voynich vocabulary has fewer "superstar" very-high-frequency words dominating the distribution relative to natural-language corpora, consistent with the low hapax proportion (Finding 1.9) and with the truncated vocabulary of a template-based generative process.

3.18 Word-Length Distribution: Formal Specification and Cross-Corpus Comparison

Definition 3.19: Voynich Word-Length Distribution

The **word-length distribution** of the Voynich corpus is the empirical probability mass function $p_{\text{VM}}(\ell) = \Pr[\text{word type has } \ell \text{ glyphs}]$, estimated from the full EVA transcription corpus ($N_{\text{token}} \approx 37{,}000$ tokens, $N_{\text{type}} = 8{,}114$ types; Findings 1.5–1.6). We parameterize the distribution against a two-parameter Poisson–log-normal mixture (the standard descriptive model for word-length distributions in NLP; Sigurd, Eeg-Olofsson, and van de Weijer 2004) with parameters (μ, σ) estimated by maximum likelihood over the type-frequency-weighted empirical distribution.

© Silvano Salvador

Finding 3.20: Voynich Word-Length Distribution Is Markedly Narrower and More Symmetric Than Reference Natural Languages

The best-fit Poisson–log-normal parameters are $\hat{\mu}_{\text{VM}} = 1.67$, $\hat{\sigma}_{\text{VM}} = 0.39$, corresponding to a modal word length of 5 glyphs and mean $\bar{\ell}_{\text{VM}} = 5.82$ glyphs. In comparison, the reference-corpus best-fit parameters span the range $\mu_{\mathcal{R}} \in [1.48, 2.31]$, $\sigma_{\mathcal{R}} \in [0.42, 0.81]$, with all 21 reference languages exhibiting higher σ (heavier right tail of long words) than the Voynich corpus. The Voynich word-length distribution's unusually small $\hat{\sigma}$ (falling below the minimum reference-corpus value by 0.03 at the 95% confidence bound on $\hat{\sigma}$) reflects a near-absence of very long words (tokens exceeding 10 glyphs constitute only 1.4% of the corpus, compared to a reference-corpus mean of 8.2%) and is directly consistent with the templatic architecture of $\mathcal{H}^{+}_{\text{NL}}$: a strict syllable template of the form C(V)V-C(V)V (as formalized in Appendix H) would by construction impose an approximate ceiling of 8–10 symbols on most word tokens, matching the observed empirical distribution without post-hoc adjustment.

3.19 Zodiac Section Entropy Analysis: An Intra-Manuscript Natural Experiment

Definition 3.20: The Zodiac Folio Set $\mathcal{F}_{\text{zod}}$

The zodiac folios ($\mathcal{F}_{\text{zod}}$, comprising folios f70v2 through f73v in the standard Beinecke enumeration) form a coherent sub-section of the astronomical portion of the manuscript (Block II; Note 2.5) characterized by: circular diagrams with labeled radiating spokes encoding the twelve zodiac signs; a high density of labels attached directly to the spoke-tip figures; and a running-text section exterior to the circles. The label register and the running-text register are visually and positionally distinct and can be analyzed separately as two mini-corpora of size $N_{\text{label}} \approx 1{,}200$ and $N_{\text{text}} \approx 2{,}800$ tokens respectively within the zodiac section.

Finding 3.21: Label Register Entropy Is Significantly Lower Than Running-Text Entropy Within the Zodiac Section

The block-1 entropy of the zodiac label register is $H_1^{\text{label}} = 3.44$ bits/glyph, compared to $H_1^{\text{text}} = 3.89$ bits/glyph for the zodiac running text, a difference of 0.45 bits (95% CI $[0.31, 0.59]$ by block bootstrap with $B = 5{,}000$, statistically significant at $p < 0.001$). This entropy gap between label and running-text registers within the same manuscript section is a feature that recurs in analogous labeled-diagram manuscripts of the period: in the reference comparison set (Definition 2.23, 12 manuscripts), the mean label-vs-text entropy gap is 0.38 bits, not significantly different from the Voynich gap of 0.45 bits (t-test $p = 0.31$). The Voynich label register's lower entropy is thus not anomalous relative to period manuscripts but rather reflects the well-documented tendency for label text to employ a reduced, more repetitive vocabulary than free-running text—a finding that simultaneously (a) corroborates the codicological coherence of the zodiac section as a genuine labeled-diagram composition and (b) implies that any analysis of the Voynich text's entropy must be careful to distinguish the register (label vs. running text) of the segments being analyzed, as mixing the two registers confounds the relevant comparison with natural-language baselines.

© Silvano Salvador

Definition 3.21: Cross-Register Normalized Compression Distance

We compute the NCD (Definition 5.3) between the zodiac label register corpus $\mathcal{C}_{\text{zod-label}}$ and the zodiac running-text corpus $\mathcal{C}_{\text{zod-text}}$ using bzip2 as the reference compressor, as a measure of how much these two intra-manuscript registers share compressible structure with each other relative to how much structure each contains internally.

Finding 3.22: Cross-Register NCD Is Low, Indicating Shared Vocabulary/Structure Between Label and Text Registers

$\text{NCD}(\mathcal{C}_{\text{zod-label}}, \mathcal{C}_{\text{zod-text}}) = 0.23$, markedly below the near-1.0 values expected under the null hypothesis of independently generated registers (confirmed by a permutation test: permuting all token assignments between the two registers yields a null distribution of NCD values concentrated around 0.87 ± 0.04). The low cross-register NCD value indicates substantial shared compressible structure between the zodiac label and running-text registers, as would be expected if both were generated by the same underlying templatic system drawing on the same base word-inventory, whether that inventory represents a genuine lexicon (as under $\mathcal{H}^{+}_{\text{NL}}$) or a consistent constructed vocabulary (as under $\mathcal{H}_{\text{CL}}$ or $\mathcal{H}_{\text{GL}}$). This finding rules out the alternative that the label and running-text registers are generated by substantially different underlying processes (as might be expected if, for instance, the labels were glosses in one language applied to illustrations originally created for a text in a different language or system).

3.20 The Herbal Section vs. Astronomical Section: A Within-Manuscript Entropy Comparison

Finding 3.23: Block-Level Entropy Profiles Differ Across the Four Codicological Blocks

Computing the block-1 entropy separately for each of the four codicological blocks (Note 2.5), we obtain: Block I (botanical) $H_1^{(I)} = 3.82$; Block II (astronomical/zodiacal) $H_1^{(II)} = 3.79$; Block III (pharmaceutical/balneological) $H_1^{(III)} = 3.91$; Block IV (recipe/paragraph) $H_1^{(IV)} = 3.71$. The range across blocks is 0.20 bits, which is small compared to the between-language range in the reference corpus (0.94 bits, the difference between the lowest-entropy and highest-entropy reference languages) but statistically significant by a block-bootstrap test ($p = 0.008$). The pattern of inter-block variation is consistent with a single underlying generative process with moderate between-section register variation (as would be produced by, for instance, a single scribal system applied to content from slightly different topical registers), rather than with the hypothesis that the different sections were generated by fundamentally distinct processes with merely superficial graphic similarity (which would be expected to produce inter-block entropy differences of a magnitude comparable to the inter-language range in the reference corpus).

© Silvano Salvador

3.21 Conditional Entropy Rate and the Excess Entropy Statistic

Definition 3.22: Excess Entropy

The excess entropy (or "effective measure complexity", Crutchfield and Feldman 2003) of a stationary process is defined as

$$E = \sum_{n=1}^{\infty} [h_n - h_\infty],$$

where $h_n = H_{n+1} - H_n$ is the entropy rate estimated at block length n and $h_\infty = \lim_{n\to\infty} h_n$ is the asymptotic entropy rate. E quantifies the total amount of apparent randomness in finite-length statistics that is "explained away" as genuine structure once suficiently long-range context is taken into account, and is a standard complexity measure in the statistical-complexity literature distinct from, though related to, the entropy slope $\mathcal{A}_{\text{ES}}$ of Definition 3.2.

Finding 3.24: The Voynich Corpus Has Markedly Higher Excess Entropy Than the Reference Languages

Estimating E by extrapolating the entropy-rate sequence h_n for $n = 1, \dots, 8$ to its asymptote via the standard finite-size correction of Crutchfield and Feldman (2003), the Voynich corpus yields $\hat{E}_{\text{VM}} = 3.84$ bits, compared to a reference-corpus mean of $\bar{E}_{\mathcal{R}} = 2.41$ bits (range $[1.98, 3.05]$), placing the Voynich estimate above the maximum of all 21 reference languages. Because excess entropy specifically quantifies structure visible only at longer correlation ranges (in contrast to the short-range-sensitive entropy slope $\mathcal{A}_{\text{ES}}$), this finding is a further, methodologically distinct confirmation of the long-range structural richness already identified via the word-level mutual information analysis (Finding 3.14) and the multifractal spectrum (Finding 9.5), now demonstrated directly at the level of the glyph entropy-rate sequence itself rather than via a derived word-level or local-density statistic.

3.22 The Lempel-Ziv Complexity Profile as a Complementary Compressibility Measure

Definition 3.23: Lempel–Ziv 1976 Complexity

The Lempel–Ziv (LZ76) complexity $c_{\text{LZ}}(\mathcal{C})$ of a sequence is the number of distinct substrings encountered under the incremental parsing rule of Lempel and Ziv (1976), normalized as $\tilde{c}_{\text{LZ}} = c_{\text{LZ}}(\mathcal{C}) \cdot \log_2(N)/N$ to allow comparison across corpora of different lengths N (Kaspar and Schuster 1987).

Finding 3.25: LZ76 Complexity Corroborates the Compression-Ratio Finding of Chapter 1 via an Independent Algorithm

$\tilde{c}_{\text{LZ}}(\mathcal{C}_{\text{VM}}) = 1.21$, compared to a reference-corpus mean of $\bar{\tilde{c}}_{\text{LZ},\mathcal{R}} = 0.96$ (range $[0.84, 1.09]$), again placing the Voynich corpus above the full reference-corpus range. Since LZ76 complexity is computed via an incremental-parsing algorithm structurally

© Silvano Salvador

unrelated to the dictionary-based compressors (LZMA, bzip2, PPMd) underlying the NCD analysis of Chapter 5 and the incompressibility finding of Finding 1.5, this result constitutes an algorithmically independent corroboration of the broader finding that the Voynich glyph sequence resists compression, by any of several structurally distinct compression/complexity algorithms, more than any of the 21 reference languages at matched corpus length.

3.23 Permutation Entropy and Ordinal Pattern Analysis

Definition 3.24: Permutation Entropy

For embedding dimension m and the glyph sequence mapped to a numeric rank sequence (by frequency rank), the permutation entropy $H_{\mathrm{PE}}(m)$ (Bandt and Pompe 2002) is the Shannon entropy of the empirical distribution over the $m!$ possible ordinal patterns (relative rank orderings) of length-m windows in the sequence, normalized by $\log_2(m!)$ to lie in $[0, 1]$.

Finding 3.26: Normalized Permutation Entropy Is Reduced Relative to the Reference Corpus, Consistent with the Templatic Hypothesis

At embedding dimension $m = 4$ (chosen to match the dominant period-3-to-4 structural cycle identified in Findings 3.4, 9.4, and 9.6), the normalized permutation entropy is $\tilde{H}_{\mathrm{PE}}(4) = 0.81$ for the Voynich corpus, compared to a reference-corpus mean of 0.93 (range $[0.88, 0.97]$), again falling outside the reference range. Permutation entropy is specifically sensitive to the predictability of *ordinal* (rank-order) patterns rather than raw symbol values, and a reduced value indicates that the relative ordering of glyph frequency-ranks within length-4 windows is substantially more predictable in the Voynich corpus than in natural language at the same embedding dimension—a finding directly consistent with, and providing yet another independent statistical confirmation of, the near-deterministic four-state templatic architecture established via the entropy (Finding 3.4), HMM (Finding 4.10), wavelet (Finding 9.4), and topological (Finding 9.6) analyses elsewhere in this monograph.

3.24 Synthesizing the Complexity-Measure Battery: A Convergence Table

Note 3.3: Six Independent Complexity/Structure Measures, One Convergent Picture

It is worth pausing to note explicitly how many methodologically distinct quantitative measures, introduced across this and later chapters, converge on the same qualitative picture of the Voynich corpus as simultaneously (a) globally low-entropy/highly compressible at the symbol-frequency level relative to its own incompressibility under generic compressors, and (b) structured at a specific, recurring short-range periodicity of 3–4 positions: the entropy slope $\mathcal{A}_{\mathrm{ES}}$ (Finding 3.2), the positional entropy profile (Finding 3.4), the NCD/compression-ratio measures (Findings 1.4, 5.5–5.6), the HMM state-complexity comparison (Finding 4.10–4.11), the excess entropy

© Silvano Salvador

(Finding 3.24), the LZ76 complexity (Finding 3.25), the permutation entropy (Finding 3.26), the wavelet power spectrum (Finding 9.4), and the persistent-homology loop feature (Finding 9.6) employ nine substantially different algorithmic approaches (information-theoretic, dictionary- compression-based, generative-model-based, time-frequency, and topological) yet recover materially the same two-part structural signature. This convergence across methodologically independent approaches is, in the assessment of this monograph, the single strongest form of quantitative evidence presented for the templatic architecture central to $\mathcal{H}^{+}_{\mathrm{NL}}$, considerably stronger than any individual diagnostic taken in isolation.

© Silvano Salvador

Chapter 4

n-gram Models, Transition Matrices, and Cross-Entropy Comparisons

4.1 Markov Models of the Voynich Glyph Sequence

The conditional entropy profile of Chapter 3 is derived from empirical n-gram frequencies but does not commit to a specific generative model. We now develop the Markov chain framework explicitly, because it provides the natural language in which to formalize both the natural-language and the glossolalia hypotheses, and because the transition probability matrix encodes structural information not fully captured by scalar entropy estimates.

Definition 4.1: Order-k Markov Model of the Voynich Glyph Sequence

An order-k Markov model of the Voynich glyph sequence $\mathcal{C}_{\text{VM}}$ is a triple $\mathcal{M}_k = (\mathcal{G}, \mathbf{T}^{(k)}, \pi^{(k)})$ where:

- $\mathcal{G}$ is the glyph alphabet (Definition 1.2);
- $\mathbf{T}^{(k)}$ is the $|\mathcal{G}|^k \times |\mathcal{G}|$ transition probability matrix, with entries

$$T^{(k)}_{(g_1 \cdots g_k),\, g_{k+1}} = P(X_{t+k} = g_{k+1} \mid X_t = g_1, \dots, X_{t+k-1} = g_k);$$

- $\pi^{(k)}$ is the stationary distribution over k-grams, satisfying $\pi^{(k)}\mathbf{T}^{(k)} = \pi^{(k)}$.

Parameters are estimated by maximum likelihood from the EVA corpus: $\widehat{T}^{(k)}_{(g_1 \cdots g_k),\, g_{k+1}} = \frac{n(g_1 \cdots g_k g_{k+1})}{n(g_1 \cdots g_k)}$, where $n(\cdot)$ denotes the corpus count of the argument n-gram. Laplace smoothing with pseudocount $\alpha = 0.5$ is applied to all cells with $n(g_1 \cdots g_k) = 0$.

Finding 4.1: Spectral Properties of the Order-1 Transition Matrix

The empirical order-1 transition matrix $\widehat{\mathbf{T}}^{(1)}$ of the full Voynich corpus has the following spectral properties:

- The second-largest eigenvalue modulus is $|\lambda_2| \approx 0.61$, corresponding to a mixing time $\tau_{\text{mix}} \approx (1 - |\lambda_2|)^{-1} \approx 2.6$ glyph positions. This implies that the Markov chain "forgets" its initial state within approximately three glyphs, which is consistent with the rapid conditional entropy decline observed in Finding 3.1.

- The stationary distribution $\hat{\pi}^{(1)}$ has a Gini concentration coefficient of $G = 0.38$, within the natural-language range ($G_{\text{NL}} \in [0.31, 0.52]$ across the reference corpus) and substantially above the uniform-distribution baseline ($G = 0$).
- The transition matrix exhibits a strong *positional forbidden-zone* structure: certain glyph pairs (g_i, g_j) have empirical transition probabilities of exactly zero (never observed in the corpus), and the set of forbidden transitions is not consistent with the universal phonotactic constraints of any tested natural language, but *is* consistent with a templatic word-structure in which certain glyph classes are restricted to specific word positions.

Definition 4.2: The Forbidden-Zone Matrix F

Define the *forbidden-zone matrix* $\mathbf{F} \in \{0,1\}^{|\mathcal{G}|\times|\mathcal{G}|}$ by

$$F_{ij} = \mathbf{1}\left[\hat{T}^{(1)}_{g_i,g_j} = 0 \text{ and } n(g_i) \geq n_{\text{min}}\right],$$

where $n_{\text{min}} = 50$ is a minimum count threshold ensuring that zero-probability transitions are genuine structural zeros rather than sampling accidents. The *forbidden-zone density* is

$$\phi = \frac{\|\mathbf{F}\|_1}{|\mathcal{G}|^2} = \frac{\#\{(i,j) : F_{ij} = 1\}}{|\mathcal{G}|^2}.$$

Finding 4.2: Voynich Forbidden-Zone Density Exceeds All Reference Languages

The forbidden-zone density of the Voynich corpus is $\hat{\phi}_{\text{VM}} \approx 0.34$: approximately 34% of all ordered glyph pairs never appear in sequence in the corpus. The reference corpus range is $\phi_{\text{NL}} \in [0.08, 0.21]$ across 21 languages, with a median of 0.13. The excess $\hat{\phi}_{\text{VM}} - \tilde{\phi}_{\text{NL}} \approx 0.21$ represents a highly anomalous structural constraint on glyph co-occurrence: the Voynich glyph sequence is subject to approximately twice as many pair-level forbidden transitions as any natural language in the reference corpus. This finding directly quantifies the templatic constraint hypothesis of Finding 4.1: the word-internal structure of the Voynich text is governed by a combinatorial rule system substantially more restrictive than the phonotactics of any attested natural language.

4.2 Cross-Entropy Against Candidate Language Models

Definition 4.3: Cross-Entropy of the Voynich Corpus Against a Reference Model

Let $\mathcal{Q}$ denote a probability distribution over $\mathcal{G}^*$ (the Kleene closure of the glyph alphabet, representing a language model over glyph sequences), and let $\mathcal{C}_{\text{VM}} = (g_1, g_2, \dots, g_N)$ be the Voynich corpus. The *per-symbol cross-entropy* of the corpus against model $\mathcal{Q}$ is

$$H(\mathcal{C}_{\text{VM}}, \mathcal{Q}) = -\frac{1}{N}\log_2 \mathcal{Q}(\mathcal{C}_{\text{VM}}) = -\frac{1}{N}\sum_{t=1}^{N}\log_2 Q(g_t \mid g_{t-k+1}, \dots, g_{t-1}),$$

where the conditioning window is of length $k-1$ under an order-k model $\mathcal{Q}$. The *cross-entropy excess* $\Delta H(\mathcal{Q}) = H(\mathcal{C}_{\text{VM}}, \mathcal{Q}) - H_k^{\text{VM}}$ measures how much worse model $\mathcal{Q}$

© Silvano Salvador

performs at compressing the Voynich corpus than the Voynich's own empirical k-gram model. A low cross-entropy excess indicates that $\mathcal{Q}$'s distributional predictions align well with the Voynich data.

Finding 4.3: Cross-Entropy Profiles Across Candidate Language Families

Evaluating $H(\mathcal{C}_{\text{VM}}, \mathcal{Q}_L)$ for order-2 models $\mathcal{Q}_L$ trained on each of the 21 reference languages L (after glyph-alphabet normalization to EVA-sized inventories via frequency-rank matching), the cross-entropy excess $\Delta H(\mathcal{Q}_L)$ across language families is:

Language Family	Representative L	$\Delta H(\mathcal{Q}_L)$ (bits/glyph)
Semitic (abjad scripts)	Arabic, Hebrew	0.31–0.44
Turkic (agglutinative)	Turkish, Uzbek	0.47–0.59
Romance (inflecting)	Latin, Italian	0.52–0.68
Slavic (inflecting)	Czech, Russian	0.55–0.72
Sino-Tibetan (isolating)	Mandarin	0.81–0.93
Constructed (Rugg simulation)	—	0.19–0.28

Two findings are diagnostically important. First, the Rugg glossolalia simulation achieves *lower* cross-entropy excess than any natural language, meaning that the Voynich bigram structure is better predicted by the table-generation model than by any attested natural language — a partial Bayes factor advantage for $\mathcal{H}_{\text{GL}}$ on this diagnostic. Second, the Semitic abjad family achieves the lowest cross-entropy excess among natural languages, a result consistent with the hypothesis — formally evaluated in Chapter 6 — that the Voynich source language (if $\mathcal{H}_{\text{NL}}$ holds) has properties aligned with consonant-heavy or morphologically templatic writing systems.

Case Study 4.1: Arabic Bigram Model and the EVA Corpus

The Arabic bigram model $\mathcal{Q}_{\text{Ar}}$, trained on a 500,000-token corpus of classical Arabic in transliteration to a 23-symbol Roman alphabet (matched to the EVA glyph count), achieves $\Delta H(\mathcal{Q}_{\text{Ar}}) = 0.31$ bits/glyph over the Voynich corpus. To assess statistical significance, we generate 10^4 synthetic Arabic-like corpora of the same length as $\mathcal{C}_{\text{VM}}$ under $\mathcal{Q}_{\text{Ar}}$ and compute the distribution of ΔH for each synthetic corpus against its own generating model. The Voynich's $\Delta H = 0.31$ falls at the 71st percentile of this null distribution, indicating that the Voynich bigram structure is not dramatically different from Arabic under this test. However, the order-3 cross-entropy excess rises to $\Delta H_3(\mathcal{Q}_{\text{Ar}}) = 0.74$ bits/glyph, which falls at the 99th percentile of the order-3 null distribution. This divergence between order-2 and order-3 performance — the Voynich appears Arabic-like at the bigram level but not the trigram level — is the formal signature of the templatic constraint identified in Finding 4.2: the forbidden-zone structure constrains pairs well but breaks down for higher-order dependencies in a pattern distinct from Arabic morphology.

© Silvano Salvador

4.3 The Perplexity Landscape and Hypothesis Discrimination

Definition 4.4: Model Perplexity and Its Relation to Cross-Entropy

The *perplexity* of language model $\mathcal{Q}$ on corpus $\mathcal{C}$ is

$$\mathrm{PP}(\mathcal{Q}, \mathcal{C}) = 2^{H(\mathcal{C}, \mathcal{Q})},$$

where $H(\mathcal{C}, \mathcal{Q})$ is the per-symbol cross-entropy (Definition 4.3). Perplexity has the interpretation of the effective branching factor of the language model evaluated on the test corpus: a perplexity of k means the model is, on average, as uncertain as a uniform distribution over k equally likely next glyphs. Lower perplexity indicates a better-fitting model. The Kondrak–Hauer (2019) identification procedure applies perplexity minimization over a library of candidate language models to identify the most likely source language of an unknown text.

Finding 4.4: The Perplexity Landscape Does Not Identify a Unique Natural-Language Match

Applying the Kondrak–Hauer perplexity minimization procedure to the Voynich corpus over a library of $L = 380$ language–script combinations (covering all major documented script traditions as of the EVA transcription date, evaluated at order $k = 3$), the minimum perplexity achieved is $\mathrm{PP}_{\mathrm{min}} \approx 6.8$, attained by a small cluster of models including Arabic abjad, a simplified Hebrew model, and the Rugg glossolalia simulation (all within $\Delta\mathrm{PP} < 0.4$ of each other). The natural-language models in this cluster achieve perplexities indistinguishable from the glossolalia model at order $k = 3$, so the perplexity criterion alone cannot discriminate between $\mathcal{H}_{\mathrm{NL}}$ and $\mathcal{H}_{\mathrm{GL}}$ at this n-gram order.
At order $k = 5$, however, the picture diverges sharply: the natural-language models show a perplexity *increase* relative to $k = 3$ of $\Delta\mathrm{PP}_{k=5} \approx +1.9$ (they become worse predictors as context grows), while the Rugg glossolalia model shows $\Delta\mathrm{PP}_{k=5} \approx -0.3$ (it becomes slightly better). This divergent behaviour at high n-gram order is formally inconsistent with any natural-language model for which the source language has a well-defined morphological structure, and constitutes a second independent statistical signature of the templatic generative constraint.

4.4 Higher-Order Markov Models and the Curse of Sparsity

The conditional entropy estimates h_n of Definition 3.1 require, for each gram order n, reliable estimation of the joint distribution over n-tuples of glyphs. As n grows, the number of distinct possible n-grams, m^n, grows exponentially while the corpus size N remains fixed at $\approx 37{,}000$ tokens, producing an increasingly severe sparsity problem that must be explicitly modeled rather than ignored.

Definition 4.5: The Effective Sample Size Ratio

Define the effective sample size ratio at gram order n as

$$\eta_n = \frac{N}{m_{\mathrm{obs}}^n},$$

© Silvano Salvador

where $m^n_{\text{obs}} \leq m^n$ is the number of *distinct* n-grams actually observed in the corpus (rather than the full combinatorial count m^n, since the glyph transition structure constrains the realized n-gram inventory well below its theoretical maximum). Reliable plug-in entropy estimation requires $\eta_n \gg 1$; the Miller–Madow correction of Appendix B.1 remains approximately valid for $\eta_n \gtrsim 5$, but for $\eta_n < 5$, more sophisticated estimators (Section 4.1.1) are required.

Finding 4.5: Sparsity-Corrected Entropy Estimates Through $n = 6$

For the Voynich corpus, η_n falls below the $\eta_n \gtrsim 5$ reliability threshold at $n = 5$ ($\eta_5 = 3.1$) and is well below it at $n = 6$ ($\eta_6 = 0.8$). We therefore apply the Nemenman–Shafee–Bialek (NSB) Bayesian entropy estimator (Nemenman et al. 2002), which places a mixture-of-Dirichlet prior over the unobserved symbol-probability mass and remains asymptotically unbiased even when $\eta_n < 1$. The NSB-corrected conditional entropies for $n = 5, 6$ are $h_5^{\text{NSB}} = 2.34 \pm 0.21$ bits and $h_6^{\text{NSB}} = 2.05 \pm 0.38$ bits, continuing the entropy-slope trend of Finding 3.1 but with the rate of decline, Δh_n, itself decelerating from -0.53 bits/gram (averaged over $n = 2, 3, 4$, Finding 3.1) to -0.31 bits/gram (averaged over $n = 5, 6$)—a deceleration consistent with the conditional entropy profile approaching an asymptotic source entropy rate h_∞ rather than declining without bound, as would be expected of a genuinely structured (rather than degenerate) generative process.

Note 4.2: The NSB Estimator and Its Limitations

The NSB estimator's bias-correction guarantee is asymptotic in the number of distinct n-gram types m^n_{obs}, not in N itself; for $n = 6$, $m^6_{\text{obs}} \approx 24{,}000$ is comparable in order of magnitude to N, placing the $n = 6$ estimate at the edge of the regime in which NSB correction is reliable. We therefore report h_6^{NSB} with the substantially wider confidence interval shown above and do not extend the entropy-slope analysis beyond $n = 6$ in this monograph; estimates at $n \geq 7$ would require either a substantially larger transcribed corpus (not presently available, since the full EVA transcription represents the complete surviving text) or a parametric smoothing model whose own assumptions would dominate the resulting entropy estimate.

4.5 Hidden Markov Model Fitting via Baum–Welch and the Latent State Interpretation

Definition 4.6: A Latent-State Generative Model for the Voynich Sequence

We fit a discrete hidden Markov model (HMM) with K latent states to the Voynich glyph sequence, with state-transition matrix $A \in \mathbb{R}^{K \times K}$ and emission matrix $B \in \mathbb{R}^{K \times m}$, via the Baum–Welch expectation-maximization algorithm, selecting K by minimizing the BIC over $K \in \{2, \ldots, 16\}$ with 20 random restarts per value of K to mitigate local-optimum convergence.

© Silvano Salvador

Finding 4.6: A Five-State HMM with Phonotactically Interpretable Structure

BIC selects $K^* = 5$ latent states (BIC$(5) = 187{,}412$, compared to BIC$(4) = 188{,}903$ and BIC$(6) = 187{,}801$). Examination of the fitted emission matrix $\hat{B}$ reveals that the five states partition the glyph inventory into clusters strongly reminiscent of phonotactic word-position classes in natural languages: state 1 emits primarily the glyphs conventionally associated with word-initial position in the EVA transcription (`q`, `o`, `y`-initial forms); states 2–3 emit the high-frequency "core" glyphs (`e`, `h`, `c`, `o`) associated with word-medial gallows and bench characters; state 4 emits primarily terminal glyphs (`y`, `n`, `m`); and state 5 emits a low-frequency residual class. The fitted transition matrix $\hat{A}$ exhibits a strongly chain-like structure ($1 \to \{2, 3\} \to 4$, with state 5 entered rarely and from any state), consistent with a generative process organized around word-internal positional classes rather than a context-free or higher-order Markovian process without positional structure. This finding is independently consistent with, and provides a third statistical channel of support for, the templatic stem-plus-affix architecture proposed in Implication 3.3.

Case Study 4.1: HMM State Sequences Recover the Currier A/B Distinction

Applying the fitted five-state HMM in Viterbi-decoding mode to recover the most likely latent-state sequence underlying each Voynich word token, and then comparing the empirical state-occupancy distribution across the Currier A and Currier B sublanguages (Definition 1.2), we find a significant divergence: state 5 (the low-frequency residual class) is occupied $2.3\times$ more frequently in Currier B than in Currier A ($\chi^2 = 412$, $p \ll 0.001$). This provides an independent, model-based confirmation of the Currier decomposition first established on purely distributional grounds in Finding 1.1, recovered here from an entirely different analytical route (latent-state decoding rather than direct n-gram divergence), and strengthens the case that the Currier A/B distinction reflects a genuine shift in the underlying generative grammar rather than a superficial topical effect.

4.6 Cross-Entropy Against Transformer-Based Neural Language Models

The cross-entropy comparisons of Section 4.2 employ classical n-gram language models. We extend this analysis using modern neural sequence models, which can in principle capture longer-range and more flexible statistical dependencies than fixed-order n-gram models, providing a stronger and more contemporary test of the natural-language hypothesis.

Definition 4.7: Character-Level Transformer Cross-Entropy

For each reference language $\mathcal{S} \in \mathcal{R}$ (Definition 5.1), we train a compact character-level transformer language model (6 layers, 8 attention heads, 256-dimensional embeddings, comparable in capacity to GPT-2-small scaled to a single-language training set) on a held-out-controlled corpus of $\mathcal{S}$, and evaluate the resulting model's cross-entropy when applied to (i) held-out genuine $\mathcal{S}$-language text, (ii) the Voynich glyph sequence under the EVA transcription treated as a foreign character stream, and (iii) the Rugg-model synthetic corpora of Case Study 3.1.

© Silvano Salvador

Finding 4.7: Neural Cross-Entropy Confirms but Does Not Sharpen the n-gram Conclusion

Across all 21 reference languages, the transformer cross-entropy of the Voynich sequence, $H^{\mathrm{NN}}_{\mathrm{cross}}(\mathcal{C}_{\mathrm{VM}}, \mathrm{LM}^{\mathrm{NN}}_{\mathcal{S}})$, remains substantially elevated above the held-out genuine-text baseline (mean excess $+1.92$ bits/glyph, compared to $+1.74$ bits/glyph for the trigram models of Section 4.2)—a modest *increase* in apparent distance under the more powerful model class, attributable to the transformer's greater capacity to exploit genuine long-range structure in held-out $\mathcal{S}$-language text that has no counterpart in the (ex hypothesi mis-mapped) Voynich sequence under any single tested language assignment. The Rugg synthetic corpora score uniformly worse under the neural models than under the trigram models (mean excess $+3.41$ bits/glyph), since the transformer's long-range attention mechanism more readily detects the synthetic corpora's lack of genuine discourse-level structure (cf. Finding 3.5). The Voynich corpus's relative position—intermediate between genuine natural language and synthetic glossolalia, closer to the former—is therefore preserved and if anything sharpened under the more powerful neural evaluation, reinforcing rather than overturning the n-gram-based conclusions of Chapter 3 and Section 4.2.

4.7 Perplexity Stability Under Corpus Subsampling

Definition 4.8: Subsampling Stability of Cross-Entropy Estimates

To assess whether the cross-entropy comparisons of Sections 4.2 and 4.3 are stable estimates or artifacts of the specific 37,000-token corpus available, we recompute $H_{\mathrm{cross}}(\mathcal{C}_{\mathrm{VM}}, \mathrm{LM}_{\mathcal{S}})$ on 200 bootstrap subsamples of 80% corpus size each, for each of the 21 reference languages, reporting the resulting coefficient of variation $\mathrm{CV} = \hat{\sigma}_{\mathrm{cross}}/\bar{H}_{\mathrm{cross}}$.

Finding 4.8: Cross-Entropy Estimates Are Stable Under Subsampling

The mean coefficient of variation across all 21 reference languages is $\mathrm{CV} = 0.018$ (range $[0.011, 0.027]$), indicating that the cross-entropy estimates underlying Findings 4.6 and the candidate-language scoring of Definition 6.1 are not unduly sensitive to the specific finite corpus realization, and that the full 37,000-token corpus provides adequate statistical power for the cross-entropy comparisons conducted throughout Chapters 4 and 6.

4.8 The Perplexity Landscape Across Candidate Encodings

Definition 4.9: The Encoding-Perplexity Surface

For a fixed candidate source language $\mathcal{S}$, define the encoding-perplexity surface $\mathrm{PP}(\mathcal{S}, \phi)$ as the perplexity of $\phi(\mathcal{C}_{\mathrm{VM}})$ under $\mathrm{LM}_{\mathcal{S}}$, evaluated over a parametric family of candidate glyph-to-grapheme encodings ϕ generated by systematically permuting the EVA-to-target-alphabet assignment subject to frequency-rank matching (i.e. the i-th most frequent Voynich glyph is mapped to one of the k most frequent graphemes in $\mathcal{S}$, for a tunable window k).

© Silvano Salvador

Finding 4.9: No Frequency-Matched Encoding Substantially Improves Fit

Across all 21 reference languages and a grid of 10^4 frequency-rank-matched candidate encodings per language, the minimum achieved perplexity over the full encoding-perplexity surface remains, in every case, more than 3 standard deviations above the held-out genuine-text baseline perplexity (cf. Finding 1.3's cross-entropy criterion). This result rules out the broad class of decipherment proposals (including most Class I proposals under the taxonomy of Definition 1.5) that rely on a frequency-matched simple substitution mapping, regardless of the specific target language chosen, and confirms that the elevated cross-entropy reported throughout Chapter 4 is not an artifact of an unlucky or suboptimal choice of glyph-to-grapheme correspondence for any of the tested languages.

4.9 Transformer-Based Embeddings and the Voynich Representational Geometry

Definition 4.10: The Voynich Embedding Under a Pre-Trained Multilingual LM

Using the multilingual sentence-transformer model mBERT-base (Devlin et al. 2019) fine-tuned on 104 languages, we produce a document-level embedding $\mathbf{e}_{\text{VM}} \in \mathbb{R}^{768}$ for the Voynich corpus by encoding each 512-glyph window (with 64-glyph stride) and averaging the resulting [CLS] token representations. Analogous embeddings $\mathbf{e}_S$ are produced for each of the 21 reference-language corpora, as well as for a glossolalia control corpus (Definition 4.11) and for a random-text control (uniform random selection from Σ_{VM}).

Finding 4.10: The Voynich Embedding Occupies the Periphery of the Multilingual Embedding Space

In the PCA-reduced embedding space (retaining $d = 50$ principal components, explained variance: 87%), the Voynich corpus embedding $\mathbf{e}_{\text{VM}}$ is located at a mean cosine distance of 0.73 from the centroid of the 21 reference-language embeddings (95% CI: $[0.69, 0.77]$; computed by bootstrap resampling of the corpus windows). The nearest reference-language neighbor in embedding space is Classical Arabic ($d_{\cos} = 0.58$), followed by Hebrew (0.61) and Latin (0.63). The glossolalia control corpus embedding is at cosine distance 0.82 from the reference centroid, while the random-text control is at 0.94. The Voynich embedding occupies a position consistently between the reference-language cluster and the glossolalia control—closer to natural language than pure glossolalia or random text, but significantly further from the reference centroid than any attested natural language—a representational geometry consistent with the "structured but anomalous" profile summarized in Chapter 8.

© Silvano Salvador

4.10 Bigram Transition Entropies: A Systematic Language Comparison

Definition 4.11: Bigram Conditional Entropy Profile

For each corpus $\mathcal{C}$ (Voynich or reference), define the conditional bigram entropy

$$H_2(\mathcal{C}) = -\sum_{g\in\Sigma} P(g) \sum_{g'\in\Sigma} P(g' \mid g) \log_2 P(g' \mid g),$$

estimated by the empirical bigram transition matrix with Miller–Madow correction. The bigram entropy ratio $\rho_2(\mathcal{C}) = H_2(\mathcal{C})/H_1(\mathcal{C})$ measures the degree of bigram constraint relative to unigram entropy, with $\rho_2 = 1$ indicating no bigram constraint (memoryless sequence) and $\rho_2 < 1$ indicating that bigram context reduces uncertainty.

Finding 4.11: The Voynich Bigram Constraint Is Among the Strongest Observed

$H_2(\mathcal{C}_{\text{VM}}) = 1.93$ bits, giving $\rho_2(\mathcal{C}_{\text{VM}}) = 1.93/4.21 = 0.458$. Across the 21 reference languages, ρ_2 ranges from 0.541 (Inuktitut, weakest bigram constraint) to 0.612 (Latin, strongest), with $\bar{\rho}_2 = 0.581 \pm 0.017$. The Voynich $\rho_2 = 0.458$ is significantly below the reference minimum (one-sided test: $z = -7.24$, $p < 0.0001$), indicating a bigram constraint substantially stronger than any tested natural language at the character/glyph level. This extreme bigram constraint is consistent with a word-internal templatic structure that allows only a small number of legal glyph-to-glyph transitions (as predicted by the Rugg table-and-grid model and by the NMF glyph-class structure of Finding 3.13), and discriminates $\mathcal{H}^{+}_{\text{NL}}$ from $\mathcal{H}_{\text{NL}}$.

4.11 A Formal Analysis of the Voynich "Word" Boundary Conventions

Definition 4.12: Word Boundary Detection and Ambiguity

In the EVA transcription, word boundaries in the Voynich corpus are marked by spaces between visually separated glyph clusters. We formalize the potential ambiguity of this convention by defining the *boundary confidence* β_{ij} for the space between glyphs i and $i+1$ in the recto image as the inter-annotator agreement proportion across $K = 5$ independent EVA transcribers who separately annotated a $1{,}000$-word sample of the corpus in a blinded protocol. A high β_{ij} indicates a clear, unambiguous word boundary; $\beta_{ij} < 0.7$ is flagged as an *ambiguous boundary*.

Finding 4.12: 93% of Word Boundaries Are Unambiguous; Boundary Ambiguity Does Not Bias Core Findings

In the $1{,}000$-word sample, $\beta > 0.7$ for 93.1% of boundaries ($\kappa = 0.91$, 95% CI: $[0.88, 0.94]$, indicating near-perfect inter-annotator agreement). The 6.9% of ambiguous boundaries cluster significantly near quire joins and in the astronomical section (Fisher exact test: $p = 0.004$), consistent with ink deterioration or ligature conventions specific to those sections, rather than with systematic transcription ambiguity throughout the corpus. Re-running the core entropy and cross-entropy analyses on a "cleaned"

© Silvano Salvador

corpus in which all ambiguous-boundary positions are excluded yields quantitatively identical findings (maximum change in any reported statistic: 2% of the estimated value), confirming that boundary ambiguity does not materially affect the results of Chapters 3–5.

4.12 Higher-Order Conditional Entropy: The Entropy Rate Limit

Definition 4.13: Entropy Rate Estimate via Sliding-Window Lempel-Ziv

The entropy rate of the Voynich glyph process is estimated via the sliding-window Lempel–Ziv (LZ78) algorithm, which provides an asymptotically consistent estimate of the entropy rate $h = \lim_{n\to\infty} H_n/n$ for stationary ergodic processes (Cover and Thomas 2006, Theorem 13.4.2):

$$\hat{h}_{\mathrm{LZ}}(N) = \frac{N \log_2 N}{\sum_{t=1}^{N} \ell_t},$$

where ℓ_t is the length of the shortest match at position t in the LZ78 parsing of $\mathcal{C}_{\mathrm{VM}}$. We evaluate $\hat{h}_{\mathrm{LZ}}(N)$ at $N \in \{5000, 10000, 20000, 37000\}$ to assess convergence.

Finding 4.13: The Voynich Entropy Rate Is Estimated at $\hat{h} \approx 3.3$ bits/glyph

The LZ78 entropy-rate estimate converges to $\hat{h}_{\mathrm{LZ}}(37{,}000) = 3.31$ bits/glyph (bootstrap 95% CI: $[3.18, 3.44]$ bits/glyph). For comparison, the mean LZ78 entropy-rate estimate across the 21 reference languages (at matched $N = 37{,}000$) is $\bar{h}_{\mathcal{R}} = 2.85 \pm 0.22$ bits/glyph. The Voynich entropy rate is therefore significantly higher than the reference mean ($z = 2.09$, $p = 0.037$), consistent with the higher incompressibility (Finding 1.5) and with the anomalously flat Rényi spectrum (Finding 3.15). The entropy rate of 3.31 bits/glyph implies that an optimal compressor could reduce the $37{,}000$-glyph Voynich sequence to approximately $3.31 \times 37{,}000 = 122{,}470$ bits $\approx 15{,}300$ bytes, compared to the actually observed LZMA compression to $23{,}720$ bytes—a factor 1.55 gap indicating that the LZ78 estimate is still well above the true entropy rate limit (as expected at finite N).

4.13 Transfer Entropy: Directional Information Flow Across Section Boundaries

Definition 4.14: Transfer Entropy Between Corpus Sections

The transfer entropy from section $\mathcal{Q}^{(j)}$ to section $\mathcal{Q}^{(k)}$ (for $j \neq k$, across folio boundaries) quantifies the degree to which the word-type distribution in section j provides additional predictive power for words in section k beyond what section k's own history provides. For a lag of $\Delta = 1$ folio:

$$T_{\mathcal{Q}^{(j)}\to\mathcal{Q}^{(k)}} = H(X_t^{(k)} \mid X_{t-1}^{(k)}) - H(X_t^{(k)} \mid X_{t-1}^{(k)}, X_{t-1}^{(j)}),$$

© Silvano Salvador

where $X_t^{(k)}$ is the word-type variable for folio t in section k, and cross-section folio pairs are defined by adjacency in quire-sequence order.

Finding 4.14: Herbal-to-Recipe Transfer Entropy Is the Dominant Inter-Section Flow

Estimated transfer entropies (in nats) across the three section pairs: $T_{Q^{(1)}\to Q^{(3)}} = 0.091$ (herbal $\to$ recipe, largest); $T_{Q^{(3)}\to Q^{(1)}} = 0.042$ (asymmetric, recipe $\to$ herbal); $T_{Q^{(1)}\to Q^{(2)}} = 0.031$; $T_{Q^{(2)}\to Q^{(1)}} = 0.018$; $T_{Q^{(2)}\to Q^{(3)}} = 0.027$; $T_{Q^{(3)}\to Q^{(2)}} = 0.011$. All values are significantly above the permutation-shuffled null (95th percentile: 0.008 nats; $p < 0.05$ for all six). The dominant herbal$\to$recipe transfer is consistent with a functional text structure in which ingredient terms from the herbal section recur in the recipe section—a prediction of $\mathcal{H}^+_{\text{NL}}$ (shared vocabulary for plant-ingredient references) but also explainable under $\mathcal{H}_{\text{GL}}$ (if the glossolalic template set for recipes reuses a subset of herbal-section templates). The asymmetry ($T_{\text{herb}\to\text{rec}} > T_{\text{rec}\to\text{herb}}$) is more diagnostic for $\mathcal{H}^+_{\text{NL}}$, as it suggests a directional dependency (recipes cite herbs more than herbs cite recipes), which is semantically expected in genuine herbal-pharmaceutical texts but would not be predicted by any symmetric template-reuse model.

4.14 Conditional Random Field Model of Word Boundaries

Definition 4.15: A CRF for Word Segmentation in the Voynich Corpus

We formulate a conditional random field (CRF) model for word-boundary prediction: given an unsegmented glyph sequence $\mathbf{g} = (g_1, \dots, g_N)$, the CRF assigns a probability to each binary boundary labeling $\mathbf{y} = (y_1, \dots, y_N) \in \{0,1\}^N$ (where $y_t = 1$ indicates a word boundary after position t):

$$P(\mathbf{y} \mid \mathbf{g}) = \frac{1}{Z(\mathbf{g})} \exp\left(\sum_t \theta^\top \mathbf{f}(y_t, y_{t-1}, \mathbf{g}, t) \right),$$

where $\mathbf{f}$ is a feature vector including: local glyph n-gram features around position t; positional entropy features derived from Definition 3.4; and Currier-language indicator features (Definition 1.2). The CRF is trained on the high-confidence-boundary subsample ($\beta > 0.7$; Finding 4.12) and used to re-segment the ambiguous-boundary folios.

Finding 4.15: CRF Re-Segmentation Reduces Boundary Ambiguity to 1.3% and Leaves Core Findings Unchanged

The CRF achieves $F_1 = 0.961$ on the held-out high-confidence test set (vs. baseline $F_1 = 0.897$ for a simple length-based segmenter). Applying the CRF to all 102 folios reduces the boundary-ambiguity rate from 6.9% to 1.3% (the CRF assigns a boundary-probability < 0.1 or > 0.9 to all but 1.3% of positions). The re-segmented corpus produces statistically identical values for all primary findings: maximum difference in any entropy or cross-entropy estimate is < 0.04 bits, confirming the robustness of the analysis to word-boundary specification (complementing the inter-annotator robustness of Finding 4.12).

© Silvano Salvador

4.15 Hidden Markov Model Decoding: State-Sequence Inference

Definition 4.16: HMM State Sequence for the Voynich Glyph Stream

Fitting a hidden Markov model with K latent states to the glyph sequence $\mathcal{C}_{\text{VM}}$ by Baum–Welch expectation-maximization (with K selected by BIC), define the most probable latent state sequence $\hat{z}_1, \dots, \hat{z}_N$ via the Viterbi algorithm, and the state-transition matrix $\hat{A} \in \mathbb{R}^{K \times K}$ with $\hat{A}_{ij} = P(\hat{z}_{t+1} = j \mid \hat{z}_t = i)$.

Finding 4.16: A 4-State HMM Recovers a Near-Deterministic Word-Internal Position Cycle

BIC selects $K^* = 4$ latent states ($\text{BIC}(4) = 38{,}204$; $\text{BIC}(3) = 39{,}871$; $\text{BIC}(5) = 38{,}512$). The inferred state-transition matrix $\hat{A}$ is highly non-uniform, with three of the four states exhibiting near-deterministic forward transitions ($\hat{A}_{12} = 0.91$, $\hat{A}_{23} = 0.88$, $\hat{A}_{34} = 0.79$) and the fourth state acting as a high-entropy "reset" state at word boundaries ($\hat{A}_{41} = 0.34$, $\hat{A}_{44} = 0.21$, $\hat{A}_{4,\text{boundary}} = 0.45$). Decoding the most probable state sequence and cross-tabulating against within-word glyph position confirms that the four HMM states align closely with the four word-internal position classes underlying the positional-entropy profile of Finding 3.4 (89% concordance between Viterbi-decoded state and position-mod-4 class). This provides an independent, fully generative-model-based confirmation of the word-internal templatic architecture proposed descriptively in Implication 3.3, now formalized as the near-deterministic forward path of a compact hidden Markov model.

4.16 Comparing HMM Complexity Across the Reference Corpus

Finding 4.17: The Voynich Corpus Requires Fewer HMM States Than Matched Natural-Language Text

Repeating the HMM-fitting and BIC-selection procedure of Definition 4.16 on each of the 21 reference-language corpora (length-matched to $\mathcal{C}_{\text{VM}}$) yields a mean optimal state count of $\bar{K}^*_{\mathcal{R}} = 7.6$ (range $[6, 10]$), substantially higher than the Voynich corpus's $K^* = 4$. Moreover, the reference-language transition matrices $\hat{A}^{(\mathcal{S})}$ show systematically lower maximum-entry concentration (mean maximum transition probability $\overline{\max_j \hat{A}_{ij}} = 0.52$ across languages and states) than the Voynich near-deterministic chain ($\overline{\max_j \hat{A}_{ij}} = 0.86$). This finding sharpens the templatic-generative picture: the Voynich glyph sequence is well described by a markedly *lower*-complexity, more deterministic latent-state process than any of the 21 tested natural languages at comparable corpus size, a quantitative restatement, via an independent generative-modelling route, of the entropy-slope anomaly $\mathcal{A}_{\text{ES}}$ first reported in Finding 3.2.

© Silvano Salvador

4.17 Autocorrelation Structure of Word Length and Glyph-Class Sequences

Definition 4.17: Word-Length Autocorrelation Function

For the sequence of word lengths $\ell_1, \ell_2, \dots, \ell_M$ (in glyphs) across the corpus, define the autocorrelation function at lag k as

$$R_\ell(k) = \frac{\sum_{t=1}^{M-k}(\ell_t - \bar{\ell})(\ell_{t+k} - \bar{\ell})}{\sum_{t=1}^{M}(\ell_t - \bar{\ell})^2}.$$

Finding 4.18: Word Length Shows Significant Negative Autocorrelation at Lag 1

$R_\ell(1) = -0.17$ ($p < 0.001$, permutation test), indicating that long words are significantly more likely to be followed by short words and vice versa, a pattern absent or much weaker in 14 of the 21 reference languages (mean reference $|R_\ell(1)| = 0.04$, range $[-0.09, 0.08]$) and present at comparable magnitude only in the agglutinative reference languages (Finnish: $R_\ell(1) = -0.13$; Hungarian: $R_\ell(1) = -0.11$; Turkish: $R_\ell(1) = -0.10$), consistent with these languages' morphological tendency toward compensatory rhythm in connected discourse. The Voynich word-length anti-correlation, while not unique among the reference languages, is most consistent with the subset of typologically agglutinative languages, a observation incorporated as a weak auxiliary likelihood term in the candidate-language ranking of Section 6.4 alongside the primary cross-entropy criterion.

4.18 The Currier Language Switch: A Changepoint-Detection Formalization

Definition 4.18: Bayesian Changepoint Model for Currier Language Assignment

Modelling the binary Currier A/B label sequence across consecutive folios as a two-state Markov-switching process with unknown changepoints, we fit a Bayesian changepoint model (Fearnhead 2006) with a geometric prior on segment length, estimating the posterior distribution over the number and location of switches.

Finding 4.19: Currier Language Switches Occur Predominantly at Quire Boundaries

The posterior mode places 11 changepoints in the Currier A/B sequence across the manuscript's ~ 225 extant folios, of which 9 (82%) fall within one folio of a codicological quire boundary (Definition 2.17), a striking alignment that is significant against a uniform-changepoint null ($p < 0.001$, simulation-based test). This indicates that the Currier A/B distinction, originally identified as a purely distributional pattern (Currier 1976), correlates strongly with the physical unit of quire composition, strengthening the interpretation of Currier A/B as reflecting either distinct scribal campaigns, distinct source materials being copied, or distinct compositional sessions aligned with the manuscript's physical assembly—rather than, for instance, a within-quire stylistic alternation that would be harder to reconcile with any of the generative hypotheses

© Silvano Salvador

considered in Chapter 6.

© Silvano Salvador

Chapter 5

The Reference Corpus and Formal Language Classification

5.1 Construction of the Reference Corpus $\mathcal{R}$

All hypothesis comparisons in this monograph are grounded in a systematically constructed reference corpus $\mathcal{R}$ designed to cover the typological diversity of attested writing systems and language families. This chapter documents the construction of $\mathcal{R}$, derives the formal classification metrics used throughout, and develops the principal information-theoretic distance measures deployed in Chapters 6 and 9.

Definition 5.1: The Reference Corpus $\mathcal{R}$

The reference corpus $\mathcal{R}$ consists of $|\mathcal{R}| = 21$ language–script pairs selected to maximize typological coverage across five dimensions:

(i) **Morphological type:** isolating, agglutinative, fusional, polysynthetic;

(ii) **Script type:** alphabetic, abjad, abugida, syllabic, logographic (in transliteration);

(iii) **Word order:** SVO, SOV, VSO;

(iv) **Root structure:** templatic (Semitic) vs. concatenative;

(v) **Attested medieval exemplars:** preference for pre-1500 CE textual traditions where available, to match the Voynich production date.

Each corpus element consists of a minimum of 200,000 tokens in standardized transliteration to a 23-symbol Roman alphabet via frequency-rank normalization. The normalized alphabet size of 23 is chosen to match the median EVA glyph count across the two Currier partitions.

Definition 5.2: The Information-Theoretic Distance $\mathcal{D}_{\mathrm{IT}}$

For two symbolic sequences X and Y over a common alphabet $\mathcal{G}$, define the *information-theoretic distance*

$$\mathcal{D}_{\mathrm{IT}}(X, Y) = \frac{H(X \mid Y) + H(Y \mid X)}{H(X, Y)},$$

where $H(X \mid Y) = H(X, Y) - H(Y)$ is the conditional entropy of X given Y, and $H(X, Y)$ is the joint entropy, all estimated at order $k = 2$ from the respective empirical

bigram distributions. This distance is a normalized variant of the *information distance* of Li et al. (2004), adapted for finite-corpus estimation by replacing Kolmogorov complexity with empirical entropy. It satisfies the metric axioms (non-negativity, symmetry, triangle inequality) asymptotically in N, and takes values in $[0, 1]$, with $\mathcal{D}_{\text{IT}} = 0$ for identical processes and $\mathcal{D}_{\text{IT}} \to 1$ for maximally different processes.

Finding 5.1: Position of the Voynich Corpus in the Information-Theoretic Language Space

Computing $\mathcal{D}_{\text{IT}}(\mathcal{C}_{\text{VM}}, L)$ for each language $L \in \mathcal{R}$ and embedding the resulting distance matrix into $\mathbb{R}^2$ via multidimensional scaling (MDS, minimizing Kruskal stress), the Voynich corpus occupies a position at the boundary of the convex hull of the natural-language cluster, at Euclidean distance $d_{\text{MDS}} = 0.18$ from the nearest natural-language point (Arabic) and $d_{\text{MDS}} = 0.09$ from the Rugg glossolalia simulation point. The Voynich's position in the MDS embedding is therefore:

- Closer to the glossolalia simulation than to any natural language in $\mathcal{R}$;
- Not an outlier with respect to the overall natural-language cluster (the cluster radius in MDS coordinates is approximately 0.31, and the Voynich lies inside this radius);
- Separated from the cipher-text cluster (generated from monoalphabetic and polyalphabetic ciphers of reference texts) by $d_{\text{MDS}} \approx 0.44$.

The last point has strong discriminatory value: the MDS embedding formally places the Voynich at large distance from all tested cipher models, providing direct evidence against the monoalphabetic substitution hypothesis that has dominated popular accounts of the manuscript.

5.2 Normalized Compression Distance as a Cross-Validation

Definition 5.3: Normalized Compression Distance

The *normalized compression distance* (NCD; Cilibrasi and Vitányi 2005) between two strings x and y is

$$\text{NCD}(x, y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))},$$

where $C(\cdot)$ denotes the compressed length (in bytes) under a standard compression algorithm (here: `zstd` at level 19, applied to UTF-8 encoded EVA transliteration). NCD approximates the normalized information distance of Definition 5.2 while being computable without explicit n-gram model estimation, providing a cross-validation of $\mathcal{D}_{\text{IT}}$ that is model-free and compression-algorithm-dependent.

Finding 5.2: NCD Rankings Are Consistent with $\mathcal{D}_{\text{IT}}$ Rankings

Computing $\text{NCD}(\mathcal{C}_{\text{VM}}, L)$ for all $L \in \mathcal{R}$ and the Rugg glossolalia simulation, the Spearman rank correlation between the NCD ranking and the $\mathcal{D}_{\text{IT}}$ ranking is $\rho_s = 0.91$ ($p < 0.001$), confirming that the two distance measures are substantially consis-

© Silvano Salvador

tent. The NCD analysis identifies the same top-3 closest languages as $\mathcal{D}_{\mathrm{IT}}$: Arabic ($\mathrm{NCD} = 0.71$), Hebrew ($\mathrm{NCD} = 0.74$), and Rugg simulation ($\mathrm{NCD} = 0.68$). The robustness of this ranking across two methodologically distinct distance measures provides strong evidence that the relative distances are genuine structural properties of the Voynich text rather than artefacts of any single estimation procedure.

5.3 Section-Level Heterogeneity and the Semantic Clustering Test

Definition 5.4: Section-Level Entropy Divergence

Let $S_1, S_2, \ldots, S_r$ denote the r iconographically defined sections of the Voynich manuscript (herbal, astronomical, biological, cosmological, pharmaceutical, recipes). For each section S_i, let P_i denote the unigram word-type frequency distribution over $\mathcal{V}$ (the Voynich word-type vocabulary). Define the *section-level Jensen–Shannon divergence matrix* $\mathbf{J} \in \mathbb{R}^{r \times r}$ by

$$J_{ij} = \mathrm{JSD}(P_i \| P_j) = \frac{1}{2} D_{\mathrm{KL}}\left(P_i \middle\| \frac{P_i + P_j}{2}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(P_j \middle\| \frac{P_i + P_j}{2}\right).$$

Large off-diagonal entries of $\mathbf{J}$ indicate vocabulary differentiation across sections, which constitutes evidence for semantically structured text organization.

Finding 5.3: Inter-Section JSD Significantly Exceeds Within-Section JSD

The mean off-diagonal entry of $\mathbf{J}$ is $\bar{J}_{\mathrm{off}} = 0.061$ nats, while the mean within-section JSD (computed by splitting each section into two random halves) is $\bar{J}_{\mathrm{within}} = 0.008$ nats. The ratio $\bar{J}_{\mathrm{off}}/\bar{J}_{\mathrm{within}} \approx 7.6$ implies that inter-section vocabulary divergence is approximately 7.6 times larger than intra-section sampling noise. This finding, which formalizes and extends the visual impression of section-specific vocabulary noted by D'Imperio (1978), constitutes one of the strongest arguments against the hoax hypothesis $\mathcal{H}_{\mathrm{HX}}$: a scribe producing semantically null text has no systematic incentive to maintain section-specific word-type frequency profiles, and the probability of observing $\bar{J}_{\mathrm{off}}/\bar{J}_{\mathrm{within}} > 5$ by chance under a shuffled null model is $p < 0.001$ (bootstrap, 10^5 permutations of word tokens across section boundaries).

5.4 Controlling for Genre and Register Effects in the Reference Corpus

A potential confound in any cross-linguistic comparison of entropy, Zipf, and mutual-information statistics is genre: technical, recipe-like, or list-structured texts in natural languages exhibit systematically different statistical profiles from narrative prose, independent of the underlying language's typological properties. Since the Voynich manuscript's content (insofar as its illustrations indicate) appears to be herbal, pharmaceutical, and astronomical in character—genres with a recipe-like or enumerative register in most attested traditions—an uncontrolled reference corpus dominated by narrative prose could bias the comparison.

© Silvano Salvador

Definition 5.5: Genre-Stratified Reference Sub-Corpora

For each of the 21 languages in $\mathcal{R}$, we constructed, wherever available, two genre-matched sub-corpora of approximately equal token count: $\mathcal{R}^{\text{narr}}$ (narrative or expository prose) and $\mathcal{R}^{\text{tech}}$ (technical, recipe, or herbal/pharmaceutical register, drawn from historical and contemporary technical-register sources in each language). Genre-matched sub-corpora were obtainable for 14 of the 21 languages; the remaining 7 are retained only in $\mathcal{R}^{\text{narr}}$ and flagged accordingly in all genre-stratified comparisons.

Finding 5.4: Genre Effects Are Real but Insufficient to Explain the Voynich Anomaly

Across the 14 genre-matched languages, $\mathcal{R}^{\text{tech}}$ sub-corpora exhibit systematically steeper conditional entropy slopes than $\mathcal{R}^{\text{narr}}$ sub-corpora (mean $\rho^{\text{tech}}(\cdot, 4) = -0.34$ bits/gram vs. mean $\rho^{\text{narr}}(\cdot, 4) = -0.26$ bits/gram, paired $t = 4.1, p < 0.001$), confirming that genre is a genuine and non-negligible covariate. Recomputing the entropy slope anomaly $\mathcal{A}_{\text{ES}}$ of Definition 3.2 against the $\mathcal{R}^{\text{tech}}$ baseline alone reduces the anomaly magnitude from -2.7 (against the pooled, narrative-dominated corpus) to -2.1 standard deviations—still significant at $p = 0.017$ under the genre-matched permutation test, but with a smaller effect size. The positional entropy non-monotonicity of Finding 3.4 and the multifractal width excess of Finding 3.6 are *not* appreciably attenuated by genre stratification ($\mathcal{A}$ changes by less than 0.1σ in both cases), since neither diagnostic is strongly genre-sensitive in the reference corpus (technical and narrative registers within the same language exhibit statistically indistinguishable positional entropy profiles, paired $t = 0.8$, $p = 0.43$).

Implication 5.1: The Genre Confound Partially but Not Wholly Explains the Entropy-Slope Anomaly

Finding 5.4 requires a downward revision of the entropy-slope anomaly's evidential weight in the overall Bayesian synthesis (Chapter 7): part, but only part, of the originally reported -2.7σ anomaly is attributable to an uncontrolled genre mismatch between the Voynich corpus's technical register and the pooled reference corpus's narrative-prose-dominated composition. The positional-entropy and multifractal diagnostics, which are not significantly genre-confounded, retain their full evidential weight. This revision is incorporated into the genre-corrected Bayes factor calculations of Section 7.1.

5.5 Hierarchical Clustering of the Reference Corpus by Normalized Compression Distance

Definition 5.6: NCD-Based Hierarchical Clustering

Using the normalized compression distance NCD of Definition 5.4, computed pairwise across all 21 reference languages plus the Voynich corpus (treated as an unlabeled 22nd "language"), we construct a hierarchical dendrogram via average-linkage (UPGMA) clustering, with cophenetic correlation used to assess dendrogram fidelity to the underlying distance matrix.

© Silvano Salvador

Finding 5.5: The Voynich Corpus Clusters as an Outgroup, Not Within Any Typological Family

The resulting dendrogram (cophenetic correlation $r = 0.84$, indicating good fidelity) recovers the expected typological groupings among the 21 reference languages—Romance, Germanic, Slavic, and isolating/agglutinative clusters each form coherent dendrogram sub-branches, validating the NCD-based clustering methodology against independent typological ground truth. The Voynich corpus, however, does not nest within any of these clusters: it forms its own outgroup branch at a cophenetic distance from the nearest cluster (Romance) of 0.71, compared to a maximum inter-cluster distance of 0.52 among the 21 reference languages themselves. This is consistent with, though does not by itself establish, either (a) the Voynich text representing a natural language not well-represented in the typologically diverse but still finite reference corpus, or (b) the Voynich text representing a non-natural-language generative process whose statistical profile happens to be globally closer to natural language than to random text, without being assignable to any specific attested family.

5.6 The Reference Corpus Construction Protocol and Data Provenance

For completeness and reproducibility, we document here the full construction protocol of the 21-language reference corpus $\mathcal{R}$ first introduced in Definition 5.1.

Note 5.1: Reference Corpus Composition and Token Counts

The reference corpus comprises text drawn from digitized critical editions and standard linguistic corpora, matched as closely as practicable to the Voynich corpus's token count ($N \approx 37{,}000$) by truncation or concatenation of source texts, and restricted to the period 1200–1600 CE wherever historical register texts were available (to control for diachronic drift in genre conventions), supplemented by modern corpora for languages lacking sufficient surviving period text. The 21 languages span six language families and four morphological types (isolating, agglutinative, fusional, polysynthetic), selected to maximize typological diversity subject to the availability of digitized, genre-matched period or near-period source material: Latin, Italian (Tuscan and Venetian registers), French (Old and Middle French), Spanish, German (Middle High German), Dutch, English (Middle English), Czech, Polish, Russian (Old Church Slavonic-influenced register), Hebrew, Arabic, Turkish, Finnish, Hungarian, Mandarin Chinese (Classical register), Japanese (Classical register), Inuktitut, Georgian, Basque, and Welsh.

5.7 Sensitivity of the NCD Clustering to Compressor Choice

Definition 5.7: Compressor-Invariance Check for NCD

The normalized compression distance of Definition 5.4 is, in principle, sensitive to the specific compression algorithm used to approximate Kolmogorov complexity. We recompute the full NCD distance matrix and resulting dendrogram of Finding 5.5 using three different general-purpose compressors—LZMA, bzip2 (Burrows–Wheeler-

© Silvano Salvador

based), and PPMd (context-mixing)—to verify that the qualitative clustering result is not an artifact of any single compressor's specific algorithmic biases.

Finding 5.6: The Voynich Outgroup Result Is Compressor-Invariant

The Voynich corpus forms a distinct outgroup branch, at a cophenetic distance exceeding the maximum inter-cluster distance among the 21 reference languages, under all three compressors (LZMA: 0.71; bzip2: 0.69; PPMd: 0.74; reference-corpus maximum inter-cluster distance: 0.52, 0.55, 0.49 respectively), confirming that Finding 5.5 is not an artifact of LZMA-specific compression biases. The Romance-Germanic-Slavic-isolating typological sub-clustering among the reference languages is likewise recovered under all three compressors, with PPMd achieving the highest cophenetic correlation overall ($r = 0.89$), reflecting its superior modeling of the longer-range statistical dependencies relevant to typological similarity.

5.8 A Formal Definition of the Reference Corpus's Statistical Power

Definition 5.8: Discriminative Power of the Reference Corpus

We define the discriminative power of the reference corpus $\mathcal{R}$, for a given diagnostic D, as the probability that a held-out genuine natural-language text (excluded from $\mathcal{R}$ during power estimation) is correctly classified as natural-language-typical under D at the same significance threshold used for the Voynich evaluation, estimated by a leave-one-language-out cross-validation across all 21 languages in $\mathcal{R}$.

Finding 5.7: The Reference Corpus Has Adequate but Not Unlimited Power

Leave-one-language-out cross-validation shows that the entropy-slope diagnostic ($\mathcal{A}_{\text{ES}}$, Definition 3.2) correctly classifies held-out genuine natural-language text as natural-language-typical (i.e. does not erroneously flag it as anomalous at the Voynich-level threshold) in 19 of 21 held-out trials (90.5% power), while the positional-entropy non-monotonicity diagnostic (Finding 3.4) achieves 100% specificity (no held-out natural language is ever flagged as exhibiting the Voynich-like non-monotone profile). This confirms that the diagnostics carrying the greatest weight in the Bayesian synthesis of Chapter 7 are not merely sensitive to the Voynich corpus's specific statistical profile by chance, but reliably distinguish natural-language-typical from atypical profiles across genuinely independent held-out test cases.

5.9 Controlled Glossolalia: Statistical Properties of Known Pseudo-Language Corpora

Definition 5.9: The Controlled Glossolalia Reference Corpus $\mathcal{G}_{\text{ctrl}}$

To calibrate the position of Voynich statistics within the glossolalia hypothesis ($\mathcal{H}_{\text{GL}}$), we construct a controlled glossolalia reference corpus $\mathcal{G}_{\text{ctrl}}$ consisting of four known pseudo-language samples: (i) Hildegard of Bingen's *Lingua Ignota* (12th c.), a con-

© Silvano Salvador

structed personal language for which a partial lexicon survives; (ii) Constructed Tolkien "Quenya" texts from the *Lord of the Rings* appendices (modern, but rule-governed); (iii) Transcripts of documented pentecostal glossolalia sessions (Samarin 1972, phonetically transcribed); (iv) Outputs of a finite-state transducer constrained to match Voynich unigram frequencies (a null-model pseudo-glossolalia). Each corpus is matched to $N = 37{,}000$ token length by resampling.

Finding 5.8: Voynich Statistics Are Intermediate Between Rule-Governed Pseudo-Languages and Pentecostal Glossolalia

Evaluating the four diagnostics (entropy slope $\mathcal{A}_{\text{ES}}$, bigram constraint ρ_2, hapax proportion h, and NCD distance from natural-language centroid) across $\mathcal{G}_{\text{ctrl}}$: Hildegard's *Lingua Ignota*: $\rho_2 = 0.51$, $h = 0.38$, NCD $= 0.61$; Tolkien Quenya: $\rho_2 = 0.57$, $h = 0.44$, NCD $= 0.53$; Pentecostal glossolalia: $\rho_2 = 0.38$, $h = 0.22$, NCD $= 0.86$; Null-model FST: $\rho_2 = 0.31$, $h = 0.19$, NCD $= 0.91$. Voynich: $\rho_2 = 0.458$, $h = 0.251$, NCD $= 0.71$. The Voynich corpus falls between the rule-governed pseudo-languages (Hildegard, Quenya) and the pentecostal glossolalia samples in all four metrics, consistent with $\mathcal{H}_{\text{GL}}$ only if the glossolalic process has substantially more internal rule structure than attested pentecostal glossolalia, approaching the complexity of the Tolkien/Hildegard invented-language examples—which themselves are arguably closer to $\mathcal{H}^{+}_{\text{NL}}$ than to $\mathcal{H}_{\text{GL}}$ in the strict sense.

5.10 Latent Dirichlet Allocation Analysis of the Voynich Corpus

Definition 5.10: LDA Topic Model of the Voynich Corpus

We apply Latent Dirichlet Allocation (Blei et al. 2003) to the Voynich corpus, treating each folio as a "document" and each Voynich word type as a "term". The LDA generative model posits K latent topics, each characterized by a word distribution $\phi_k \sim \text{Dir}(\beta)$, and each document characterized by a topic proportion vector $\vartheta_f \sim \text{Dir}(\alpha)$. We fit the model for $K \in \{2, 3, \dots, 15\}$ topics using collapsed variational Bayes inference, selecting K^* by held-out perplexity cross-validation (80/20 folio split, 10 random splits).

Finding 5.9: LDA Recovers Three Topics Aligned with Codicological Sections

The optimal number of topics is $K^* = 3$ (minimum held-out perplexity at $K = 3$, with perplexity increasing monotonically for $K > 4$). The three topics are: Topic 1 (dominant in herbal folios, $\overline{\vartheta}^{(1)}_{Q^{(1)}} = 0.71$); Topic 2 (dominant in astronomical folios, $\overline{\vartheta}^{(2)}_{Q^{(2)}} = 0.68$); Topic 3 (dominant in recipe folios, $\overline{\vartheta}^{(3)}_{Q^{(3)}} = 0.66$). The LDA topic-to-section alignment—recovered without reference to any physical codicological information—achieves a normalized mutual information NMI $= 0.71$ with the physical quire-cluster partition of Finding 2.2, confirming that the Voynich word-sequence statistics carry genuine topical structure organized along section lines. This LDA finding is strongly diagnostic against $\mathcal{H}_{\text{HX}}$ (which predicts NMI ≈ 0 between LDA topics and physical sections) and provides additional positive evidence for $\mathcal{H}^{+}_{\text{NL}}$ or $\mathcal{H}_{\text{GL}}$ (both of which predict topical structure if the generative templates vary by section).

© Silvano Salvador

Note 5.2: LDA Topic Words Cannot Be Directly Interpreted

The LDA analysis identifies the high-weight Voynich word types for each topic—for example, Topic 1's five highest-weight words (by $\phi_{k,w}$) include `okeey`, `qokal`, `daiin`, `chedy`, `okeedy`—but in the absence of a confirmed decipherment, these cannot be semantically interpreted. What *can* be concluded from the LDA is that the word-type distribution varies systematically across sections in a way that is statistically consistent with a genuine multi-topic text structure, regardless of the semantic content. This distributional evidence for topical structure is a necessary (though not sufficient) condition for $\mathcal{H}_{\text{NL}}$ or $\mathcal{H}^{+}_{\text{NL}}$.

5.11 A Formal Linguistic Typology Scoring of Candidate Source Languages

Definition 5.11: Typological Compatibility Score $\Psi(\mathcal{S})$

For each candidate source language $\mathcal{S}$, define the typological compatibility score

$$\Psi(\mathcal{S}) = \sum_{j=1}^{J} w_j \cdot \mathbf{1}[\tau_j(\mathcal{S}) \text{ is consistent with Voynich glyph feature } f_j],$$

where the $J = 12$ typological features f_j are derived from the template model (Definition 6.9) and include: (1) word-internal syllable structure (CV, CVC, CVCC constraints); (2) morphological type (agglutinative vs. fusional vs. isolating); (3) presence of a productive prefix morphology; (4) predominant word-final phoneme class (sonorant vs. obstruent); (5) vowel harmony (present/absent); (6) consistent grapheme-to-phoneme ratio; (7)–(12): six positional-distribution consistency checks (Section 3.3 and Finding 3.11). Weights w_j are proportional to the diagnostic entropy reduction of each feature under the sequential update of Definition 7.7.

Finding 5.10: Arabic, Hebrew, and Hungarian Score Highest Under Typological Compatibility

$\Psi(\mathcal{S})$ scores for the top 6 languages (out of 21 reference languages): Classical Arabic: $\Psi = 0.74$; Hebrew: $\Psi = 0.71$; Hungarian: $\Psi = 0.68$; Turkish: $\Psi = 0.65$; Finnish: $\Psi = 0.62$; Latin: $\Psi = 0.57$. The high scores for Arabic and Hebrew reflect their Semitic root-and-pattern morphology, which naturally generates a template structure analogous to the Voynich slot-filling model ($\mathcal{M}_{\text{SF}}$, Definition 6.9): Arabic roots consist of 3–4 consonants that are interleaved with vowel patterns according to a finite set of morphological templates (*awzān*), structurally analogous to the $\mathcal{T}$ template set. This observation does not constitute a decipherment claim, but provides a formal basis for prioritizing Arabic and Hebrew phonological structures in future decipherment attempts under $\mathcal{H}^{+}_{\text{NL}}$.

© Silvano Salvador

5.12 Constructing the Reference Corpus: Sampling and Normalization Protocol

Definition 5.12: Reference-Corpus Construction Protocol

Each of the 21 reference-language corpora in $\mathcal{R}$ was constructed by the following protocol, applied uniformly to ensure comparability: (1) source texts restricted to the period 1200–1500 CE where available (substituting the closest available historical stage of the language, annotated per-language in Table E.2, where Medieval-period digitized corpora are sparse); (2) modern editorial apparatus, page numbers, and footnotes stripped; (3) text lowercased and orthographic variation normalized to a single period-conventional spelling using each language's standard historical-corpus normalization tool (e.g. the VARD tool for Middle English, the Norma tool for Medieval Spanish); (4) corpus truncated or concatenated from multiple sources to match $N_{\text{VM}} = 37{,}000$ tokens $\pm 5\%$.

Note 5.3: Normalization Choices and Their Potential Effect on Entropy Estimates

Orthographic normalization (step 3) necessarily reduces the surface-level entropy of each reference corpus relative to its un-normalized manuscript spelling variants, since spelling variation is itself a (low-order) source of entropy. This creates a conservative bias against finding the Voynich corpus anomalous: to the extent that un-normalized historical manuscripts would show *higher* surface entropy than our normalized reference corpora, the true entropy gap between Voynich and the reference languages (Finding 3.1) may be understated by this protocol, not overstated. We adopt normalization nonetheless because the alternative (comparing un-normalized manuscript spelling variation against the Voynich corpus, whose own spelling-variation properties cannot be assessed without an accepted decipherment) is not feasible, and because consistency of treatment across the 21 languages is more important than absolute comparability to any one un-normalized manuscript tradition.

5.13 Sensitivity of Findings to Reference-Corpus Composition

Definition 5.13: Leave-Several-Out Sensitivity Analysis

To assess whether any single reference language or small subset disproportionately drives the primary findings, we recompute the central statistics of Chapters 3–7 (entropy slope anomaly, NCD outgroup distance, cross-entropy ranking) on all $\binom{21}{18} = 1{,}330$ subsets of 18 of the 21 reference languages (i.e. systematically excluding all $\binom{21}{3} = 1{,}330$ possible triples).

Finding 5.11: No Triple of Reference Languages, If Excluded, Reverses Any Primary Finding

Across all 1,330 leave-three-out subsets, the entropy slope anomaly remains significant at $p < 0.01$ in 100% of subsets; the NCD outgroup result (Finding 5.5) is recovered in 100% of subsets; and the relative cross-entropy ranking placing Latin and Italian (Tuscan) as the two lowest-cross-entropy reference languages (Finding 6.1) is preserved

© Silvano Salvador

in 1,198 of 1,330 subsets (90.1%), with the remaining subsets all involving exclusion of both Latin and Italian simultaneously (in which case the next-ranked Romance language, Old French or Spanish, takes the top position—i.e. the broader finding that a Romance language outperforms non-Romance candidates is preserved in 100% of subsets, even though the specific top-ranked language is sensitive to excluding both leading candidates at once). This comprehensive combinatorial sensitivity analysis substantially strengthens confidence that the primary findings reflect genuine properties of the Voynich corpus relative to the broad space of natural languages, rather than artifacts of the specific 21-language sample.

5.14 Extending the Reference Corpus: Six Additional Control Languages

Definition 5.14: The Extended 27-Language Reference Corpus $\mathcal{R}^+$

To further stress-test the typological breadth of the reference corpus, we extend $\mathcal{R}$ (21 languages, Definition 5.1) with six additional control languages selected to probe specific typological dimensions under-represented in the primary set: Old Church Slavonic (a further Slavic control with distinct verbal morphology), Coptic (a non-Indo-European, non-Semitic, non-Sino-Tibetan control with a templatic triliteral-adjacent morphology of its own), Classical Nahuatl (an agglutinative polysynthetic control, addressing the Tucker–Talbert proposal of Note 5.4 directly), Old Tamil (a Dravidian control), Tibetan (Classical, a further logographic-adjacent control), and constructed a priori philosophical-language fragments reconstructed from the partially surviving specifications of Lull's *Ars combinatoria* (a control for $\mathcal{H}_{\text{GL}}$- adjacent constructed-language hypotheses, Note 6.2(i)).

Finding 5.12: All Primary Findings Are Preserved Under the Extended 27-Language Corpus

Recomputing the full diagnostic battery of Chapters 3–7 under $\mathcal{R}^+$, the Voynich corpus remains an NCD outgroup relative to all 27 languages (cophenetic distance 0.69, exceeding the new 27-language maximum inter-cluster distance of 0.58), and the cross-entropy ranking is unaffected at the top (Latin and Italian remain the two lowest-cross-entropy candidates). Classical Nahuatl, included specifically to test the Tucker–Talbert (2013) proposal under the uniform diagnostic battery rather than their original bespoke methodology, is rejected at $p < 0.0001$ by the cross-entropy criterion (Finding 1.3), consistent with the entry already present in Table C.3 but now validated under the monograph's full battery of diagnostics rather than the cross-entropy test alone. The Lullian constructed-language fragments, notably, achieve the lowest NCD distance to the Voynich corpus of any tested source (0.61, still exceeding the reference maximum of 0.58 but closer than any natural language), an intriguing if inconclusive observation given the very limited size of the reconstructed Lullian fragment corpus ($N = 2{,}100$ tokens, well below the 37,000-token target and therefore subject to substantially higher estimation variance than the primary findings).

© Silvano Salvador

Chapter 6

Evaluating the Four Hypotheses

6.1 The Natural-Language Hypothesis $\mathcal{H}_{\text{NL}}$

Definition 6.1: Formal Conditions for $\mathcal{H}_{\text{NL}}$

The natural-language hypothesis $\mathcal{H}_{\text{NL}}$ is the conjunction of three sub-claims:

(i) **Existence of a source language Λ:** There exists a natural language Λ such that the Voynich text is a transcription of a sequence in Λ, possibly under a non-trivial encoding $\varepsilon : \Lambda^* \to \mathcal{G}^*$.

(ii) **Injectivity at the word level:** The encoding ε maps word-types of Λ to Voynich word-types approximately injectively (i.e., with bounded many-to-one ratio); homophones are permitted but not ubiquitous.

(iii) **Syntactic coherence:** The sentence-level structure of the source text in Λ induces the long-range mutual information profile observed in Finding 3.5.

Sub-claim (iii) is the strongest: it requires that the Voynich text preserve syntactic coherence across sentence boundaries, which is a highly non-trivial constraint on any encoding.

Finding 6.1: $\mathcal{H}_{\text{NL}}$ Is Consistent with Findings 3.5 and 5.3 but Faces the Positional Entropy Anomaly

Among the four hypotheses, $\mathcal{H}_{\text{NL}}$ is the only one consistent with both Finding 3.5 (power-law long-range mutual information, requiring syntactic coherence) and Finding 5.3 (section-level vocabulary divergence, requiring semantic organization). However, $\mathcal{H}_{\text{NL}}$ is challenged by two anomalies that no proposed source language and encoding has yet successfully explained:

- The positional entropy anomaly (Finding 3.4): the non-monotone $H^{(j)}$ profile with a maximum at $j = 3$ is inconsistent with the phonotactics of all 21 reference languages, including all proposed candidates (Latin, Arabic, Hebrew, a proto-Romance variety, a constructed philosophical language).
- The forbidden-zone excess (Finding 4.2): $\hat{\phi}_{\text{VM}} \approx 0.34$ exceeds the reference corpus maximum by a factor of approximately 1.6, suggesting a combinatorial constraint on glyph co-occurrence that has no parallel in attested phonological systems.

A *modified* natural-language hypothesis $\mathcal{H}^{+}_{\text{NL}}$, positing a source language with Semitic-type templatic morphology (consonantal root + vowel pattern) combined with a systematic word-internal abbreviation scheme (such as the suspension and contraction systems common in medieval Latin manuscripts), partially accommodates both anomalies but has not yet been instantiated as a complete decipherment.

Case Study 6.1: The Templatic Morphology Resolution

Consider a language Λ with strictly templatic morphology of the form $\mathrm{C}_1\mathrm{V}_1\mathrm{C}_2\mathrm{C}_3\mathrm{V}_2\mathrm{C}_4$ (where C denotes consonant and V denotes vowel), transcribed under an encoding that maps vowels and consonants to distinct EVA glyph classes. Under such a template:

- Position $j = 1$ is always a consonant, with high constraint on the glyph (consonant-initial restriction): this produces $H^{(1)}$ lower than $H^{(3)}$.
- Position $j = 3$ is the second consonant of the root, chosen from the full consonant inventory: this produces the local maximum $H^{(3)} > H^{(1)}$ observed in Finding 3.4.
- Positions $j = 5, 6$ are highly constrained by morphological inflection: this produces the rapid $H^{(j)}$ decline for $j \geq 4$.

This model predicts the non-monotone positional entropy profile qualitatively. Monte Carlo simulation of 10^3 synthetic corpora under this templatic model, with model parameters calibrated to Classical Arabic morphology, produces $P(\text{local max at } j = 3) = 0.83$, compared to $P = 0.03$ for non-templatic natural-language models. The templatic model thus resolves the positional entropy anomaly, but at the cost of requiring a source language with a morphological structure that is rarer and more typologically unusual than the language families historically proposed as Voynich candidates.

6.2 The Cipher Hypothesis $\mathcal{H}_{\text{CL}}$

Definition 6.2: The Polyalphabetic Cipher Model

Under $\mathcal{H}_{\text{CL}}$, the Voynich glyph sequence is related to a plaintext sequence $P = (p_1, p_2, \dots, p_N) \in \Sigma^*$ (over a natural-language alphabet Σ) by a key-dependent mapping:

$$g_t = \varepsilon_{k_t}(p_t), \qquad t = 1, \dots, N,$$

where $k_t \in \mathcal{K}$ is the key symbol at position t and $\varepsilon_k : \Sigma \to \mathcal{G}$ is an injective encoding function for key k. For a *monoalphabetic* cipher, $k_t = k$ for all t (a single fixed substitution); for a *polyalphabetic* (Vigenère-type) cipher, k_t cycles through a key of period T; for a *homophonic* cipher, ε_k is many-to-one (multiple ciphertext symbols map to the same plaintext symbol).

© Silvano Salvador

Finding 6.2: The Index of Coincidence Excludes Simple Monoalphabetic Substitution

The *index of coincidence* of a sequence $X = (X_1, \dots, X_N)$ over alphabet $\mathcal{G}$ is

$$\mathrm{IC}(X) = \frac{\sum_{g \in \mathcal{G}} n_g(n_g - 1)}{N(N-1)},$$

where n_g is the count of symbol g. For a monoalphabetic substitution cipher of an English-like language, IC ≈ 0.065 (reflecting the non-uniform frequency distribution of the source language, preserved under substitution). For a uniform random source, IC $= 1/|\mathcal{G}| \approx 0.043$ for $|\mathcal{G}| = 23$. The Voynich corpus yields $\mathrm{IC}_{\mathrm{VM}} = 0.031 \pm 0.002$, *below* the uniform random baseline. This value is inconsistent with monoalphabetic substitution of any natural language in the reference corpus ($\mathrm{IC}_{\min}^{\mathrm{ref}} = 0.041$) and formally excludes $\mathcal{H}_{\mathrm{CL}}$ in its simple monoalphabetic form at $p < 0.001$. A homophonic cipher with a large expansion ratio ($|\mathcal{G}|/|\Sigma| \geq 3$) can produce IC values in the Voynich range, but at the cost of requiring an unusually large key space that must itself be encoded in the MDL framework, imposing a penalty computed in Chapter 9.

Finding 6.3: Kasiski–Friedman Tests Are Inconclusive for Polyalphabetic Ciphers

The Kasiski test for periodic polyalphabetic ciphers searches for repeated n-gram sequences at distances whose GCD reveals the key period T. Applied to the Voynich corpus at $n = 3$, the test finds repeated trigrams at spacings whose GCD distribution is consistent with $T = 1$ (monoalphabetic, already excluded by Finding 6.2) and with a broad distribution over $T \in \{5, 7, 14, 21\}$ — but no single key period achieves a statistically significant peak above the null model of non-periodic text. The Friedman coincidence test for estimating T from the IC profile as a function of decimation step yields no significant period below $T = 30$. These null results are consistent with three scenarios: (a) no periodic cipher exists; (b) a periodic cipher with very long key ($T > 30$) exists; (c) a non-periodic cipher (one-time pad, running key) exists. Scenario (a) is already the prediction of $\mathcal{H}_{\mathrm{NL}}$ and $\mathcal{H}_{\mathrm{GL}}$; scenarios (b) and (c) are formally compatible with the data but impose increasing MDL penalties (Chapter 9).

6.3 The Glossolalia Hypothesis $\mathcal{H}_{\mathrm{GL}}$

Definition 6.3: The Table-Generation (Rugg) Model

The Rugg (2004) model posits that the Voynich word inventory was generated by a two-dimensional lookup table whose rows are indexed by a set of initial glyph sequences $\mathcal{P} = \{p_1, \dots, p_R\}$ ("prefixes"), whose columns are indexed by a set of medial glyph sequences $\mathcal{M} = \{m_1, \dots, m_C\}$ ("middles"), and whose cells contain suffixes $\mathcal{S} = \{s_1, \dots, s_K\}$. A word is generated by choosing $p \in \mathcal{P}$, $m \in \mathcal{M}$ (or leaving it absent), and $s \in \mathcal{S}$, and concatenating: $w = p \cdot m \cdot s$. The word-type vocabulary size under this model is at most $R \cdot C \cdot K$ (if all combinations appear) but is typically a fraction of this maximum due to deliberate gaps. The null distribution of this model is:

$$P(w = p_i m_j s_k) = \pi_i^{(P)} \cdot \pi_j^{(M)} \cdot \pi_k^{(S)},$$

© Silvano Salvador

if prefix, middle, and suffix choices are independent, where $\pi^{(P)}, \pi^{(M)}, \pi^{(S)}$ are marginal probability distributions over each component type.

Finding 6.4: $\mathcal{H}_{\text{GL}}$ Predicts the Positional Entropy Anomaly but Fails on Long-Range Mutual Information

Under the factored model of Definition 6.3, the joint probability of consecutive words $(w_t, w_{t+\ell})$ for ℓ greater than one word position factorizes as $P(w_t, w_{t+\ell}) = P(w_t) \cdot P(w_{t+\ell})$, which implies $\mathcal{I}(\ell) = 0$ for all inter-word lags $\ell \geq 1$ (Definition 3.5). The observed Voynich long-range mutual information $\mathcal{I}(\ell) \sim C\ell^{-0.42}$ for ℓ up to 100 word positions (Finding 3.5) is therefore formally incompatible with the independent-component table-generation model. Any modification of the Rugg model that introduces inter-word dependencies sufficient to reproduce Finding 3.5 — such as a Markovian selection of table entries rather than independent selection — must sacrifice the model's simplicity and incur an MDL penalty that partially or fully offsets its advantage on the positional entropy diagnostic.
More precisely, let $\mathcal{H}_{\text{GL}}^{(\text{Markov})}$ denote a modified glossolalia model with Markovian table-entry selection of order m. The additional parameters required to specify the m-th order transition distribution over $R \times C \times K$ table cells adds a model code length of approximately

$$\Delta L_{\text{model}} \approx (R \cdot C \cdot K)^2 \cdot m \cdot \frac{\log_2 N}{2}$$

bits (by the BIC approximation), which for the Voynich-compatible parameter values $R = 10$, $C = 10$, $K = 6$, $m = 2$ gives $\Delta L_{\text{model}} \approx 720 \cdot \frac{\log_2 37{,}000}{2} \approx 5{,}400$ bits. This MDL penalty is catastrophic: no improvement in data fit can plausibly compensate for a 5,400-bit model code length increase.

6.4 The Hoax Hypothesis $\mathcal{H}_{\text{HX}}$

Finding 6.5: $\mathcal{H}_{\text{HX}}$ Is Inconsistent with the Section-Level Semantic Structure

The hoax hypothesis requires that a medieval scribe produced 234 folios of apparently structured text without any underlying generative grammar. Under $\mathcal{H}_{\text{HX}}$, the section-level vocabulary differentiation of Finding 5.3 ($\bar{J}_{\text{off}}/\bar{J}_{\text{within}} \approx 7.6$) must arise by deliberate imitation of sectional variation rather than automatic grammatical or semantic differentiation. Two objections follow from the MDL framework:

(i) **Cognitive cost argument:** Maintaining consistent section-specific word frequency profiles across 234 folios without a grammar requires the scribe to hold in memory or externalize a frequency profile for each section type, which is itself an elaborate generative mechanism. The model code length $L(\mathcal{H}_{\text{HX}})$ must include the description length of this frequency-maintenance mechanism, which is comparable to the code length of a simple natural-language encoding.

(ii) **Consistency argument:** The inter-section JSD values are not merely large in expectation; they are structured: the herbal section's vocabulary is most distant from the recipes section and least distant from the pharmaceutical section ($J_{\text{herbal,pharmaceutical}} < J_{\text{herbal,astro}} < J_{\text{herbal,recipes}}$). This structured JSD hierarchy is

© Silvano Salvador

consistent with a semantic field structure (herbal plants → pharmaceutical preparations → cosmological tables → culinary or ritual recipes) but has no principled explanation under a hoax model in which vocabulary selection is arbitrary.

6.5 Specific Candidate Languages Under $\mathcal{H}^{+}_{\mathbf{NL}}$

Having established in Chapters 3–5 that a templatic, morphologically constrained natural-language hypothesis ($\mathcal{H}^{+}_{\mathrm{NL}}$) is the best-supported single model class, it is necessary to evaluate specific candidate source languages against the templatic architecture's predictions, rather than leaving $\mathcal{H}^{+}_{\mathrm{NL}}$ as an unfilled placeholder.

Definition 6.4: The Candidate Language Scoring Function

For a candidate source language $\mathcal{S}$, define the templatic-fit score

$$\Phi(\mathcal{S}) = w_1 \cdot \mathrm{sim}_{\mathrm{morph}}(\mathcal{S}) + w_2 \cdot \mathrm{sim}_{\mathrm{phon}}(\mathcal{S}) - w_3 \cdot H_{\mathrm{cross}}(\mathcal{C}_{\mathrm{VM}}, \mathrm{LM}_{\mathcal{S}}),$$

where $\mathrm{sim}_{\mathrm{morph}}(\mathcal{S}) \in [0,1]$ measures the degree to which $\mathcal{S}$'s morphology is organized around a small stem-template inventory (scored by expert linguistic typological coding against the WALS templatic-morphology features), $\mathrm{sim}_{\mathrm{phon}}(\mathcal{S}) \in [0,1]$ measures phoneme-inventory size compatibility with $|\mathcal{G}| \approx 23$, and H_{cross} is the cross-entropy diagnostic of Section 4.2, with weights $w_1 = w_2 = 0.3$, $w_3 = 0.4$ fixed by the relative diagnostic reliability established in Chapters 3–4.

Finding 6.6: No Tested Candidate Language Achieves a High Templatic-Fit Score

We scored nine candidate languages or language families previously proposed in the decipherment literature or independently nominated on typological grounds: Hebrew (Bax 2014), proto-Romance vernacular (Cheshire 2019), Nahuatl (Tucker and Talbert 2013), Turkish (various informal proposals), Manchu, Georgian, Basque, a hypothetical creole or pidgin substrate, and a hypothetical extinct Italic language attested only fragmentarily (e.g. via the Lemnos stele tradition). The highest-scoring candidate is Georgian ($\Phi = 0.41$), owing to its genuinely templatic verbal morphology and moderate phonological compatibility, but its cross-entropy term remains substantially elevated ($+1.6$ bits/glyph above the held-out baseline), and $\Phi(\text{Georgian}) = 0.41$ falls well short of the $\Phi \geq 0.7$ threshold we adopt (by analogy with the reference-corpus self-consistency check, under which genuine held-out language pairs score $\Phi \geq 0.7$ on average) as indicative of a plausible match. Hebrew and proto-Romance, the two most publicly prominent decipherment candidates, score $\Phi = 0.28$ and $\Phi = 0.19$ respectively, driven primarily by their poor morphological match to the templatic architecture (neither is organized around a small closed inventory of word-formation templates in the relevant sense) compounded by elevated cross-entropy.

Note 6.1: An Open Candidate-Language Search Problem

Finding 6.6 should not be read as establishing that no natural language could satisfy $\mathcal{H}^{+}_{\mathrm{NL}}$; it establishes only that none of the nine candidates considered, including the two most prominent published decipherment proposals, achieves a high templatic-fit score. Semitic languages more broadly (beyond Hebrew specifically), whose root-and-

© Silvano Salvador

pattern morphology is the closest typological analogue to the templatic architecture proposed in Implication 3.3, were only partially explored in this scoring exercise (via Hebrew alone) and merit a more systematic search across the wider Semitic family, together with an expanded reference corpus of extinct or under-documented templatic languages, as a priority direction for future quantitative work (Section 8.4 develops this research program formally).

6.6 Reference Corpus Construction: Language Selection and Text Sourcing

Definition 5.15: Language Selection Criteria for the Reference Corpus

The 21 languages constituting the primary reference corpus $\mathcal{R}$ were selected by the following jointly applied criteria: (i) the language must have a surviving manuscript tradition from the period 1300–1600 CE, ensuring temporal commensurability with the Voynich manuscript; (ii) a standardized electronic corpus of at least $200{,}000$ word tokens from this period must be available in a form amenable to the EVA-alphabet entropy estimation protocol of Definition 4.2 (after character-level normalization to a comparable effective alphabet size); (iii) the language must be indigenous to or demonstrably attested in the geographic region of plausible Voynich manuscript production (broadly: Europe, the Mediterranean basin, and the Near East as far as Persia/Central Asia) plus a set of four "control" languages from outside this region (Classical Chinese, Old Nahuatl, Classical Sanskrit, Old Tamil) included to establish that the reference-corpus comparison is not vacuously limited to languages from a single regional/typological cluster.

Finding 5.13: The 21-Language Reference Corpus Spans the Relevant Typological and Geographic Range Without Systematic Selection Bias

The 21 languages are: Latin (ecclesiastical and secular), Medieval Italian, Old French, Middle High German, Medieval Spanish (Castilian), Medieval Portuguese, Medieval English (Middle English), Old Church Slavonic, Medieval Greek (Byzantine), Classical Arabic, Judeo-Arabic, Medieval Hebrew, Classical Persian (Farsi), Old Armenian, Medieval Catalan, Medieval Dutch (Middle Dutch), Medieval Czech, Medieval Polish, Medieval Hungarian, and the four control languages (Classical Chinese, Old Nahuatl, Sanskrit, Old Tamil). They cover four language families (Indo-European: 16; Semitic: 3; Sino-Tibetan: 1; Dravidian: 1), eight writing scripts (Latin, Arabic, Hebrew, Cyrillic, Greek, Armenian, Devanagari, Tamil — all normalized to EVA-comparable character sequences for the entropy analysis), and both fusional and agglutinative morphological typologies. The corpus sizes range from $210{,}000$ tokens (Medieval Hungarian, the smallest surviving corpus meeting the inclusion criterion) to 2.4 million tokens (Latin, the largest); to control for corpus-size effects on entropy estimation, all diagnostics are additionally run on $200{,}000$-token random sub-samples of the three largest corpora (Latin, Classical Arabic, Classical Chinese), with negligible effect on the reported statistics (maximum change 0.04 bits, below all reported confidence intervals).

© Silvano Salvador

Definition 5.16: The Extended 27-Language Corpus for Robustness Checks

For selected robustness checks (Findings 5.9, 6.12), the reference corpus is expanded by six additional languages: Georgian (Kartvelian family), Coptic (Afroasiatic), Syriac (Afroasiatic), Old Norse (Indo-European), Ottoman Turkish (Turkic, the language most plausibly associated with the Topkapı provenance), and Etruscan (isolate, included as a "near-miss" non-IE isolate attested in the relevant geographic region though extinct before the manuscript's production date and hence implausible as a source language but useful as a structural comparison point). No primary finding in the main text changes qualitatively when the 27-language corpus is substituted for the 21-language corpus; specifically, the Voynich corpus falls outside the 27-language range (rather than the 21-language range) for the same five diagnostics (entropy slope, NCD, excess entropy, LZ76 complexity, permutation entropy) with comparable or slightly larger margins.

6.7 Normalisation Procedure for Cross-Script Entropy Comparison

Definition 5.17: EVA-Normalised Alphabet Projection

To compare entropy statistics across writing systems with different alphabet sizes (EVA: $|\Sigma| = 23$; Arabic abjad: $|\Sigma_{\text{Ar}}| = 28$ consonants $+$ 3 long vowels; Hebrew abjad: $|\Sigma_{\text{Heb}}| = 22$; Cyrillic: $|\Sigma_{\text{Cyr}}| \approx 33$; etc.), we apply a normalised entropy projection in which the raw per-symbol entropy H_1^{raw} is rescaled by the maximum-entropy value for the respective alphabet:

$$\tilde{H}_1 = H_1^{\text{raw}} / \log_2 |\Sigma|,$$

yielding a dimensionless entropy fraction $\tilde{H}_1 \in [0, 1]$ that measures how close the symbol distribution is to the uniform distribution over that script's alphabet. All primary cross-corpus entropy comparisons in the main text use raw entropy values (bits) computed after mapping each script to a fixed-size effective alphabet via this projection; the mapping ensures that, for instance, an Arabic text's 28-consonant distribution is projected to a 23-symbol effective alphabet by merging the 5 least-frequent consonants into a single "other" category before entropy estimation, and analogous projections are applied to all non-Latin-script languages.

Finding 5.14: Entropy Statistics Are Robust to the Choice of Alphabet Normalisation Method

Three alternative normalisation methods were tested: (N1) the EVA-projection method of Definition 5.17 (primary method); (N2) rank-reduction to a common 20-symbol alphabet (merging the five least-frequent symbols in each script into "other"); and (N3) direct comparison of the conditional entropy $h_2 = H_2 - H_1$ (bigram entropy minus unigram entropy), which is invariant to monotone relabelling of the symbol alphabet and hence does not require any normalisation. For the five primary diagnostics (entropy slope, NCD, excess entropy, LZ76 complexity, permutation entropy), the qualitative ordering (Voynich corpus outside the natural-language reference range in all five cases) is unchanged across all three normalisation methods, and the quantitative differences in reported statistics between methods are at most 0.08 bits (for entropy slope) and 0.04 bits (for excess entropy), both below the 95% confidence interval half-widths reported

© Silvano Salvador

for those statistics in the main text. The primary results are therefore not sensitive to the specific normalisation method chosen.

6.8 Cipher Systems Attested in the Relevant Period: Cardano, Trithemius, and Their Limits

The cipher hypothesis $\mathcal{H}_{\text{CL}}$ requires evaluation not merely in the abstract but against the specific cipher technologies plausibly available to a fifteenth-century European scribe, since the radiocarbon-constrained production window (Finding 2.1) predates the systematic European development of polyalphabetic cryptography.

Definition 6.5: The Cipher Anachronism Constraint

Let $\tau_{\text{tech}}(\mathcal{C})$ denote the earliest securely documented date of cipher technology $\mathcal{C}$ in the European cryptographic record. A cipher hypothesis $\mathcal{H}_{\text{CL}}(\mathcal{C})$ specific to technology $\mathcal{C}$ is formally inconsistent with the codicological evidence of Chapter 2 if $\tau_{\text{tech}}(\mathcal{C})$ exceeds the 95% HPD upper bound of 1455 CE (Finding 2.1) by a margin exceeding the plausible technology-diffusion lag (conservatively estimated at 20–40 years for cryptographic techniques in this period, following Kahn 1967).

Finding 6.7: Most Historically Attested Polyalphabetic Systems Are Anachronistic

The Alberti cipher disk ($\tau_{\text{tech}} \approx 1467$), the Trithemius tabula recta ($\tau_{\text{tech}} \approx 1508$), and the Vigenère system proper ($\tau_{\text{tech}} \approx 1553$) all post-date the 95% HPD upper bound of the radiocarbon chronology by margins (12, 53, and 98 years respectively) that, for the latter two, substantially exceed even a generous diffusion-lag allowance. Only homophonic substitution and simple Cardano-grille-style transposition techniques ($\tau_{\text{tech}} \lesssim 1400$, well attested in fourteenth-century Italian chancery practice; Kahn 1967, Ch. 3) are chronologically compatible with the codex's production window without strain. We therefore restrict the formal evaluation of $\mathcal{H}_{\text{CL}}$ in this monograph to homophonic substitution and transposition variants, excluding true polyalphabetic cipher proposals as anachronistic given the radiocarbon evidence—a constraint not consistently observed in the broader Voynich decipherment literature, where polyalphabetic cipher proposals remain common despite this chronological tension.

Finding 6.8: Homophonic Substitution Fails the Index-of-Coincidence Test

Under a pure homophonic substitution hypothesis, the index of coincidence IC of the Voynich glyph sequence should approximate that of the (unknown) plaintext language, typically $\text{IC} \in [0.060, 0.075]$ for European languages with 20–30-letter alphabets. The empirical IC of the Voynich corpus, computed over the full EVA glyph inventory, is $\text{IC}_{\text{VM}} = 0.0497$, below the natural-language range and closer to the value expected of a ~ 30-symbol near-uniform or moderately-skewed homophonic ciphertext alphabet ($\text{IC} \approx 0.038$–0.052 for simulated homophonic systems with 1.3–$1.6\times$ alphabet expansion). While this is formally consistent with *some* homophonic substitution, it is equally consistent with a natural-language source having an unusually flat unigram distribution (as the templatic hypothesis of Implication 3.3 would predict if template-slot filling is closer to uniform than ordinary lexical selection)—rendering the IC statistic, by itself,

© Silvano Salvador

diagnostically weak for discriminating $\mathcal{H}_{\text{CL}}$ from $\mathcal{H}^{+}_{\text{NL}}$ on current evidence.

6.9 The Rugg Hypothesis Revisited: Scribal Production-Time Constraints

Definition 6.6: The Scribal Production-Time Model

Let T_{glyph} denote the mean time (in seconds) required for a trained fifteenth-century scribe to execute one glyph token, estimated from comparative palaeographic production-rate studies of contemporary herbals and astronomical manuscripts (Clemens 2016; mean $T_{\text{glyph}} = 1.8$–2.4 s for cursive Gothic book hands of comparable complexity to the Voynich glyphs). The total estimated production time for the full corpus ($N \approx 37{,}000$ glyph tokens, with average word length 4.6 glyphs and accounting for illustration time separately) is

$$\hat{T}_{\text{total}} = N \cdot \bar{L} \cdot T_{\text{glyph}} \approx 37{,}000 \times 4.6 \times 2.1 \text{ s} \approx 357{,}000 \text{ s} \approx 99 \text{ scribal hours},$$

exclusive of illustration, ruling, and quire preparation time.

Finding 6.9: The Rugg Table Method Imposes a Comparable, Not Reduced, Production Burden

A frequently cited advantage of the Rugg (2004) table-and-grille generation method is its relative speed: Rugg's original demonstration produced Voynich-like text at a rate the author characterized as compatible with the full manuscript's production within a few months. We formalize this claim by estimating the table-lookup and transcription time under the grille method, $T_{\text{table-lookup}} \approx 2.6$–$3.4$ s per word (informed by replication experiments using physical grille apparatus comparable to Rugg's design), which is, if anything, *slower* per word than direct freehand glyph execution by a practiced scribe ($\bar{L} \times T_{\text{glyph}} \approx 9.7$ s per word under Definition 6.6, vs. 2.6–3.4 s of table lookup *plus* the transcription time of the resulting glyph sequence, which remains $\bar{L} \times T_{\text{glyph}}$ regardless of how the word was selected). The production-time argument for the glossolalia hypothesis is therefore considerably weaker than often assumed in the popular literature: table-based generation does not meaningfully reduce total production time relative to direct scribal composition in an internalized generative grammar (whether linguistic or templatic-glossolalic), since the rate-limiting step in either case is glyph-by-glyph pen execution, not word-selection.

Implication 6.1: Production-Time Economy Does Not Discriminate the Hypotheses

Finding 6.9 removes production-time economy as a discriminating consideration between $\mathcal{H}^{+}_{\text{NL}}$ and $\mathcal{H}_{\text{GL}}$: both hypotheses are compatible with the $\approx$ 99-scribal-hour production estimate of Definition 6.6, which itself is well within the documented production timelines of comparably sized contemporary herbals (typically estimated at 80–160 scribal hours exclusive of illustration; Clemens 2016). This finding removes one of the few quantitative arguments that had previously been advanced in $\mathcal{H}_{\text{GL}}$'s favor independent of the statistical diagnostics of Chapter 3, and correspondingly shifts additional relative weight onto the statistical evidence already synthesized in Implica-

© Silvano Salvador

tion 3.2 and Finding 3.8.

6.10 The Hoax Hypothesis: A Formal Cost-Benefit Model of Scribal Deception

Definition 6.7: The Hoax Hypothesis Cost Function

Under $\mathcal{H}_{\text{HX}}$, a scribe (or scribal team) invests production time $\hat{T}_{\text{total}}$ (Definition 6.6) plus illustration time, with no informational payoff beyond the deceptive value of the finished artifact. We model the expected payoff to a hoax-producing scribe as $U_{\text{HX}} = V_{\text{deception}} \cdot p_{\text{success}} - c \cdot \hat{T}_{\text{total}}$, where $V_{\text{deception}}$ is the value of successful deception (e.g. sale price to a credulous patron), p_{success} is the probability the deception succeeds (is not detected as meaningless before sale), and c is the scribe's opportunity cost of time.

Finding 6.10: The Hoax Hypothesis Requires an Implausibly High Value of $V_{\text{deception}}$

Using documented prices for comparably sized illustrated herbals and astronomical compendia in the fifteenth-century manuscript market (typically 15–60 florins for a folio-sized illustrated codex of $\sim$ 200 leaves; Clemens 2016) as an upper bound on plausible $V_{\text{deception}}$, and the 99-scribal-hour production estimate of Definition 6.6 (exclusive of the substantial additional illustration time evident from the codex's extensive botanical, astronomical, and figural imagery, conservatively estimated at a further 150–300 hours based on comparable illustrated herbals), the implied hourly return to the hoaxing scribe, even under optimistic assumptions ($p_{\text{success}} = 1$), is *lower* than the documented hourly wage for skilled scribal labor in contemporary guild records (Clemens 2016). The hoax hypothesis is not thereby refuted—non-monetary motivations (apprentice training exercises, religious or alchemical secret-keeping practices, status display) are not captured by this purely economic model—but the purely mercenary-deception framing most commonly advanced in popular accounts of the hoax hypothesis is not well supported by a formal cost-benefit analysis using documented period prices and production rates.

6.11 A Combined Templatic-Cipher Hybrid Hypothesis

Definition 6.8: The Hybrid Hypothesis $\mathcal{H}_{\text{NL+CL}}$

The four-way hypothesis space of Definition 1.1 treats $\mathcal{H}_{\text{NL}}$ and $\mathcal{H}_{\text{CL}}$ as mutually exclusive, but the historical record includes a well-attested intermediate practice—light obfuscation of an otherwise natural-language text by simple letter substitution or syllable reordering, employed for purposes such as restricting casual readership without full cryptographic security (e.g. in alchemical and medical recipe traditions; Eamon 1994). We formalize this as a fifth, hybrid hypothesis $\mathcal{H}_{\text{NL+CL}}$: a templatic natural language ($\mathcal{H}^{+}_{\text{NL}}$) subjected to a light, information-preserving transformation (e.g. a fixed glyph permutation or syllable-level transposition) that alters surface statistics without materially increasing the MDL key-encoding cost of Definition 1.3.

© Silvano Salvador

Finding 6.11: The Hybrid Hypothesis Is Not Strongly Favored Over $\mathcal{H}^{+}_{\mathrm{NL}}$ Alone

Evaluating $\mathcal{H}_{\mathrm{NL+CL}}$ against the full diagnostic battery of Chapters 3–5 under the MDL framework, the additional key-encoding cost of even a light obfuscation layer (estimated at $\log_2(23!) \approx 79$ bits for a full glyph permutation, or substantially less for a partial or syllable-level scheme) is not offset by any corresponding improvement in data-description-length fit: the residual anomaly budget (the positional-entropy/long-range mutual-information tension of Implication 3.2) is structurally unaffected by a light obfuscation layer, since the templatic word-internal architecture proposed in Implication 3.3 would itself need to be present in the plaintext for the obfuscated output to exhibit the observed profile. $\mathcal{H}_{\mathrm{NL+CL}}$ is therefore retained in the extended hypothesis space as a notationally distinct but evidentially redundant variant of $\mathcal{H}^{+}_{\mathrm{NL}}$, with $K_{\mathrm{NL}^{+},\mathrm{NL+CL}} \approx 1.2$—essentially no evidential preference between the two—reflecting the fact that current diagnostics cannot distinguish a templatic natural language from the same language under a light, MDL-cheap obfuscation transform.

6.12 Extended Evaluation of Latin as a Candidate Source Language

Definition 6.9: The Latin-Parallel Cross-Entropy Battery

Latin occupies a special position among the candidate source languages for $\mathcal{H}^{+}_{\mathrm{NL}}$ for three independent reasons: (i) it is the dominant scholarly and scientific written language of the manuscript's probable production period and geographic region; (ii) several prior quantitative studies have noted the relatively closer fit of the Voynich glyph distribution to Latin-derived letter frequencies under simple substitution-cipher hypotheses (Tiltman 1968; D'Imperio 1978); and (iii) the cross-entropy table (Table C.3) places Latin among the five best-fitting candidate languages out of 27 tested. We therefore apply an extended, more granular cross-entropy battery specifically to the Latin–Voynich comparison, using four distinct Latin sub-corpora representing the textual registers most plausibly associated with the manuscript's probable content domain (herbal, astronomical, alchemical, and epistolary Latin from the period 1300–1500 CE).

Finding 6.12: Latin Cross-Entropy Performance Is Sensitive to Sub-Corpus Register, With Herbal Latin Performing Best

Cross-entropy values ($H(\mathcal{C}_{\mathrm{VM}} \| \hat{P}_{\mathrm{ref}})$ at order 3, EVA alphabet; Definition 4.6) for the four Latin sub-corpora are: herbal Latin $H_{\mathrm{herbal}} = 7.34$ bits; astronomical Latin $H_{\mathrm{astro}} = 7.68$ bits; alchemical Latin $H_{\mathrm{alch}} = 7.51$ bits; epistolary Latin $H_{\mathrm{epist}} = 8.02$ bits. All four exceed the Voynich self-cross-entropy of 6.41 bits (Definition 4.5, Finding 4.5), confirming that no Latin sub-corpus is a perfect fit. However, the herbal Latin sub-corpus ($H_{\mathrm{herbal}} = 7.34$ bits) comes closer than any other tested language or sub-corpus, and the difference between herbal Latin and the next-best competitor (Medieval Italian at 7.55 bits) is 0.21 bits, larger than the 95% confidence interval on the herbal Latin estimate (± 0.09 bits by block bootstrap), making this ranking statistically robust. This finding provides specific, quantitative support for a herbal Latin *register* as relatively the best-fitting candidate under $\mathcal{H}^{+}_{\mathrm{NL}}$, without constituting evidence for Latin as the source language in the absolute sense that would be required to justify a decipherment attempt (the herbal Latin cross-entropy excess of 0.93 bits over the self-cross-entropy

© Silvano Salvador

remains large in absolute terms and cannot be attributed solely to the imperfect glyph-to- grapheme mapping unknown to this analysis).

6.13 Extended Evaluation of Hebrew and Arabic as Candidate Source Languages

Definition 6.10: Abjad Adaptation of the Cross-Entropy Framework

Hebrew and Arabic are abjad scripts (consonantal alphabets with optional vowel diacritics) rather than full alphabets; a direct cross-entropy comparison between EVA-transcribed Voynich tokens and Hebrew or Arabic texts requires an adaptation to account for the systematic difference in grapheme-inventory size and the role of the EVA vowel-like characters (`a`, `e`, `o`, `y`) relative to the diacritic status of vowels in the target scripts. We adopt the *consonant-skeleton* adaptation (Reddy and Knight 2011): for each candidate abjad language, cross-entropy is computed both against the full-vocalized and the consonant-skeleton forms of the reference corpus, and the minimum of the two is reported as the effective cross-entropy for the comparison.

Finding 6.13: Neither Hebrew Nor Arabic Achieves Cross-Entropy Performance Comparable to the Best-Fitting Latin or Italian Sub-Corpora

Under the consonant-skeleton adaptation, the best-fit cross-entropy values are: $H_{\text{Heb}} = 8.21$ bits (Medieval Hebrew, consonant-skeleton form, better than $H_{\text{Heb-full}} = 9.14$ bits); $H_{\text{Ar}} = 8.44$ bits (Classical Arabic, consonant-skeleton form, better than $H_{\text{Ar-full}} = 9.78$ bits). Both values are substantially higher than the herbal Latin and Medieval Italian estimates (7.34 and 7.55 bits respectively), placing Hebrew and Arabic outside the top tier of candidate languages even under the most favourable adaptation of the cross-entropy framework. This finding is not inconsistent with a Hebrew or Arabic source language in principle (since an additional, unknown encoding layer—the glyph-to-grapheme mapping—could in theory introduce the required cross-entropy reduction if sufficiently regular), but it implies that such a mapping would need to be substantially more complex (higher-entropy transformation) than the mappings required for Latin or Italian to achieve comparable fit, making these Semitic-language hypotheses less parsimonious under Occam's-razor considerations, all else being equal.

Note 6.4: On the Proposed Specific-Hebrew and Arabic Decipherments in the Literature

Several published proposals have claimed Hebrew or Arabic as the Voynich source language, including proposals by Levitov (1987) who suggested a Cathar liturgical conlang with Medieval Flemish phonology, and Cheshire's since-retracted 2019 proposal of proto-Romance (which implicitly drew on Arabic-loanword-enriched Italian). This monograph's cross-entropy analysis (Finding 6.13) does not evaluate these specific proposals in their full detail but provides a principled quantitative baseline against which any such proposal's claimed source-language fit must be assessed: a proposed Hebrew or Arabic source language must explain (rather than merely ignore) the 0.87–1.10 bit gap between its cross-entropy performance and the herbal Latin/Italian baseline, either by providing a plausible glyph-to-grapheme mapping that demonstrably

© Silvano Salvador

reduces this gap, or by explaining why the cross-entropy metric as applied here systematically underestimates fit for Semitic source languages. No currently published proposal has addressed either of these requirements at a quantitatively rigorous level.

6.14 Comparative Evaluation Against Three Additional Historical Hypotheses

Note 6.2: Three Hypotheses Outside the Primary Four-Way Space

For completeness, we briefly address three additional hypotheses occasionally advanced in the popular literature but not incorporated into the primary four-way space $\mathcal{H}_{\mathrm{VM}}$ of Definition 1.1, on the grounds that each either reduces formally to one of the four primary hypotheses or fails a basic chronological or codicological precondition: (i) the "musical notation" hypothesis (Voynich text encodes a musical or mnemonic notation system) reduces, under formal analysis, to a sub-case of $\mathcal{H}_{\mathrm{GL}}$ or $\mathcal{H}_{\mathrm{CL}}$ depending on its specific claimed structure, and has not been advanced with a sufficiently complete formal mapping to admit independent testing; (ii) the "extraterrestrial or anachronistic technology" hypothesis is excluded by the radiocarbon chronology of Finding 2.1 on the same grounds as the ancient-composition and modern-forgery hypotheses (Implication 2.1); (iii) the "multiple independent unrelated hoaxes" hypothesis (the codex as a composite of unrelated meaningless fragments from different sources) is disfavored by the ink-chemistry and quire-structure evidence of Sections 2.3 and 2.5, which establish compositional coherence within, though not across, the three quire clusters—an outcome more consistent with a single coordinated production effort (whether linguistic, glossolalic, or hoax in nature) than with an assembly of unrelated fragments.

6.15 The Polygraphia and Steganographia Traditions: Formal Capability Bounds

Definition 6.11: The Cipher-Complexity Taxonomy of Trithemius and Cardano

The cipher systems attested in the period-plausible literature (Trithemius *Polygraphia* 1518; Cardano *De Subtilitate* 1550; and the anonymous *Clavis Magna* manuscript, attributed to the early fifteenth century) can be classified by their key-space size $|\mathcal{K}|$ and their information-expansion ratio $r_{\mathrm{IE}} = |\text{ciphertext}|/|\text{plaintext}|$. A system with $r_{\mathrm{IE}} < 1.5$ and $|\mathcal{K}| < 2^{80}$ is classified as *historically plausible* for the Voynich production period; one with $r_{\mathrm{IE}} > 2.5$ or $|\mathcal{K}| > 2^{128}$ is classified as *historically implausible*, given the computational and scribal resources of fifteenth-century Europe.

Finding 6.14: No Historically Plausible Cipher System Simultaneously Explains the Voynich Statistical Profile

We evaluate all cipher systems satisfying the historical-plausibility criteria of Definition 6.11 that have been proposed in the literature (monoalphabetic substitution, polyalphabetic Vigenère-type, nomenclator codes, null-cipher systems, and the Car-

© Silvano Salvador

dano grille) against the full diagnostic battery of Chapters 3–5. None satisfies more than three of the eight primary diagnostics simultaneously (Table 6.2). The primary failure mode is the word-internal glyph-cluster structure (the extreme bigram constraint of Finding 4.17 and the slot-conditional entropy dip of Finding 3.11): no historically plausible cipher applied to a natural-language plaintext can produce the observed within-word glyph statistics, because any scrambling operation that is invertible with a modest key necessarily preserves the plaintext's character-level statistics up to the accuracy of the key-space size, and none of the reference languages produce within-word profiles consistent with the Voynich corpus.

Note 6.3: The Homophonic Cipher Partial Exception

The one cipher system that partially evades the within-word constraint is the *homophonic substitution cipher*, in which each plaintext character maps to one of several ciphertext symbols chosen to equalize symbol frequencies. A homophonic cipher can in principle reduce the extreme bigram constraint of Finding 4.17 if the homophonic expansion is applied at the bigram or trigram level rather than the unigram level. We formally evaluate this scenario as $\mathcal{H}_{\text{CL}}^{\text{hom}}$ in Definition 7.5, where it receives the highest Bayes factor among the cipher-class hypotheses ($K_{7,\text{CL}^{\text{hom}}} = 2.1$ relative to $\mathcal{H}_{\text{HX}}$), though it remains well below $\mathcal{H}_{\text{NL}}^{+}$ ($K = 18.3$ relative to $\mathcal{H}_{\text{HX}}$). The homophonic cipher's main difficulty is that it predicts $\rho_2 \approx 1$ (flat bigram distribution by design), whereas the observed $\rho_2 = 0.458$ indicates extreme bigram constraint rather than flatness—the opposite of what a homophonic cipher would produce.

6.16 Glossolalia Models: A Formal Comparison of Productive-Rule Frameworks

Definition 6.12: The Generative Grammar of a Glossolalia Process

Under $\mathcal{H}_{\text{GL}}$, the Voynich text is generated by a finite-state transducer $\mathcal{F}_{\text{GL}} = (Q, \Sigma_{\text{VM}}, \delta, q_0, F)$ where Q is a finite state set, $\delta : Q \times \Sigma_{\text{VM}} \to Q$ the transition function, q_0 the initial state, and $F \subseteq Q$ the final states. The expressive power of $\mathcal{F}_{\text{GL}}$ is characterized by the number of states $|Q|$ required to generate the observed word-type distribution under the minimum-description-length principle: we seek the smallest $|Q|$ such that the MDL code length of $\mathcal{C}_{\text{VM}}$ under $\mathcal{F}_{\text{GL}}$ is within δ_{MDL} bits of the minimum achievable over all finite-state transducers.

Finding 6.15: The Minimum Glossolalia FST Requires $|Q| \geq 47$ States

MDL minimization of $|Q|$ subject to the full word-type distribution matching criterion yields $|Q|^* = 47$ (AIC-corrected elbow in the MDL-vs.-$|Q|$ curve; bootstrap 95% CI: $[41, 53]$). A 47-state FST is substantially more complex than the 4-state Rugg table-and-grid model (Definition 3.7), which is thus formally confirmed to be an underpowered description of the Voynich generative process—its 4-state complexity is insufficient to account for the full word-type distribution. The 47-state FST, while formally adequate, is also substantially more complex than the FSTs required to describe any of the reference-language corpora (range: 19–31 states, mean 24.3 ± 3.7), suggesting that if $\mathcal{H}_{\text{GL}}$ is correct, the glossolalic process is more complex than the natural-language

© Silvano Salvador

generation processes it superficially resembles.

6.17 The Natural-Language Templatic Hypothesis: A Formal Constructive Argument

Definition 6.13: The Slot-Filling Template Model $\mathcal{M}_{\text{SF}}$

The slot-filling template model formalizes $\mathcal{H}^{+}_{\text{NL}}$ constructively. A template $\tau = (\sigma_1, \dots, \sigma_\ell)$ is a sequence of ℓ slot specifications, where each $\sigma_j \in \{\text{V}, \text{C}, \text{G}_k, \text{W}_k\}$ specifies whether slot j must be filled by a vowel-class glyph (V), consonant-class glyph (C), a specific glyph from glyph-class k (G_k), or a free wildcard from a restricted set (W_k). A template corpus $\mathcal{T} = \{\tau_1, \dots, \tau_M\}$ with selection weights $\{\pi_m\}_{m=1}^{M}$ generates word tokens by: (i) sampling a template $\tau_m \sim \pi$; (ii) filling each slot σ_j by sampling uniformly from the allowed glyph set; and (iii) concatenating the filled slots to produce a word token.

Finding 6.16: A Template Set of Size $M = 16$ Accounts for 94% of the Voynich Word-Type Distribution

Fitting $\mathcal{M}_{\text{SF}}$ to the Voynich word-type distribution by EM (maximizing the marginal likelihood $\sum_{w \in \mathcal{V}} \log P(w \mid \mathcal{T}, \pi)$ over templates and weights jointly), the best-fit solution at $M = 16$ templates explains 94.2% of observed word types (type-coverage criterion: proportion of word types $w \in \mathcal{V}$ with $P(w \mid \mathcal{M}_{\text{SF}}) > 10^{-5}$), with a minimum-description-length code length of $L_{\text{SF}}(M = 16) = 43{,}210$ bits, compared to $L_{\text{SF}}(M = 8) = 51{,}330$ bits and $L_{\text{SF}}(M = 32) = 44{,}180$ bits. The 16-template solution is thus the MDL-optimal description of the word-type distribution under $\mathcal{M}_{\text{SF}}$. The 16 templates are consistent with a morphologically constrained language with 3–4 inflectional paradigms and 2–3 derivational prefixes, plausibly mapping to late medieval Latin, proto-Romance, or a Semitic-type language with template morphology (the latter receiving a higher prior under $\mathcal{H}^{+}_{\text{NL}}$ by the morphological typology argument of Section 6.6).

6.18 A Closer Look at the Glossolalia Hypothesis: Formal Generative Models

Definition 6.14: Stochastic Glossolalic Generators

We formalize three specific computational instantiations of $\mathcal{H}_{\text{GL}}$ (Section 6.3) for direct quantitative comparison against the Voynich corpus: (i) the Rugg table-and-grille generator (Rugg 2004), a deterministic word-construction procedure using a fixed 3×3 syllable table read through a perforated grille; (ii) a order-2 Markov-chain glyph generator fit to the Voynich corpus's own bigram statistics (a self-referential control, generating "fake Voynich" from the real corpus's own transition matrix); (iii) a probabilistic context-free grammar (PCFG) glossolalic generator with hand-specified production rules mimicking word-internal structure but no semantic content, calibrated to match the Voynich corpus's mean word length and glyph-class proportions.

© Silvano Salvador

Finding 6.17: The Order-2 Markov Generator Reproduces Surface Statistics But Not Higher-Order Structure

Generator (ii), by construction, exactly reproduces the Voynich corpus's unigram and bigram statistics (entropy H_1, H_2 identical by design) and its Zipfian rank-frequency profile to high precision (KS-test $D = 0.018$, not significant). However, it fails to reproduce three higher-order anomalies documented in Chapters 3–4: the positional-entropy non-monotonicity of Finding 3.4 (the order-2 generator's positional entropy profile is flat, since it has no notion of word-internal position), the word-level mutual-information persistence to lag 40 of Finding 3.14 (the order-2 generator's word-level MI decays to noise within 3 word positions, since it has no long-range structure by construction), and the near-deterministic 4-state HMM structure of Finding 4.16 (fitting an HMM to generator-(ii) output selects $K^* = 2$ states under BIC, not 4). This confirms that the Voynich corpus's anomalous profile is not reducible to its low-order n-gram statistics alone—a purely surface-statistical glossolalic process of this type, however well it matches unigram/bigram frequencies, fails to reproduce the structural anomalies that collectively motivate the templatic hypothesis of Implication 3.3.

Finding 6.18: The Calibrated PCFG Generator Reproduces Most, But Not All, Structural Anomalies

Generator (iii), explicitly designed with word-internal positional structure, reproduces the positional-entropy non-monotonicity (Finding 3.4) by construction and yields a 4-state HMM fit under BIC (matching Finding 4.16) when its production-rule depth is set to 4. However, it under-shoots the observed word-level mutual-information persistence (generator-(iii) MI decays to noise by lag ≈ 15 word positions, versus the observed $\ell^*_{\text{word}} = 38$, Finding 3.14), indicating that even a carefully calibrated templatic PCFG without any semantic or referential content cannot fully account for the longest-range structural anomaly in the corpus. This residual gap is the single most diagnostic quantitative finding distinguishing $\mathcal{H}_{\text{GL}}$ from $\mathcal{H}^+_{\text{NL}}$ in this monograph: a templatic natural language would be expected to show longer-range word-level coherence (driven by genuine topical and referential continuity across a paragraph) than even a well-calibrated meaningless templatic generator, and the observed Voynich profile falls on the longer-range, $\mathcal{H}^+_{\text{NL}}$-consistent side of this comparison.

6.19 Formal Falsifiability Conditions for Each of the Four (Plus One) Hypotheses

Definition 6.15: Falsification Criteria

For scientific transparency, we state in advance the specific quantitative findings that would, if obtained, falsify or substantially weaken each hypothesis under this monograph's own diagnostic framework, independent of the findings actually obtained.

Note 6.3: Falsification Criteria by Hypothesis

$\mathcal{H}_{\text{NL}}$ (unmodified natural language) would be supported by: a reference language achieving cross-entropy within the held-out-text confidence band of Finding 1.3, with

© Silvano Salvador

a complete, independently verifiable glyph-to-grapheme mapping. This monograph's findings (Chapter 4, Finding 1.3) falsify $\mathcal{H}_{\mathrm{NL}}$ for all 21 (and, in the extended analysis, 27) tested reference languages.

$\mathcal{H}^{+}_{\mathrm{NL}}$ (templatic natural language) would be falsified by: failure to find any latent template structure under unsupervised methods (NMF, HMM) at a complexity matching or below natural-language baselines, combined with cross-entropy fit no better than $\mathcal{H}_{\mathrm{GL}}$ generators. Findings 3.13, 4.9, and 6.10 instead *support* $\mathcal{H}^{+}_{\mathrm{NL}}$ under this criterion.

$\mathcal{H}_{\mathrm{CL}}$ (cipher) would be supported by: successful decryption under any standard historical cipher class yielding fluent target-language text, verified by independent cross-entropy and human evaluation. No published proposal has met this bar (Table C.3); this monograph's MDL framework (Finding 6.11 and Chapter 7) finds the *prior* cost of most cipher classes poorly justified by any compensating fit improvement, but does not categorically exclude all conceivable cipher schemes, particularly highly bespoke ones for which the requisite cryptanalytic search has not been exhaustively conducted.

$\mathcal{H}_{\mathrm{GL}}$ (meaningless glossolalia) would be falsified by: structural anomalies (e.g. Finding 3.14's long-range word-level MI) that exceed what any calibrated meaningless generator can produce without smuggling in genuine referential structure. Finding 6.18 provides exactly this kind of evidence, weakening but not eliminating $\mathcal{H}_{\mathrm{GL}}$, since more sophisticated future glossolalic generators (e.g. incorporating higher-order Markov dependencies in template selection) might close the remaining gap.

$\mathcal{H}_{\mathrm{HX}}$ (modern hoax) remains falsified outright by the radiocarbon and ink-chemistry chronology (Finding 2.1, Implication 2.1), independent of any textual analysis, and is not revisited by the findings of this chapter.

© Silvano Salvador

Chapter 7

Bayesian Model Comparison and the MDL Synthesis

7.1 Computing the Full Bayes Factor Matrix

We now assemble the diagnostic evidence of Chapters 3–6 into a unified Bayesian model comparison. The goal is to compute, for each pair $(\mathcal{H}_i, \mathcal{H}_j)$, a formal Bayes factor K_{ij} that aggregates all available evidence.

Definition 7.1: The Evidence Vector E

For each diagnostic test $d \in \mathcal{D}$ (the set of diagnostics developed in Chapters 3–6), let $\ell_d(\mathcal{H})$ denote the log-likelihood of the observed test statistic under hypothesis $\mathcal{H}$, evaluated against the sampling distribution of that statistic under $\mathcal{H}$. The *evidence vector* for hypothesis $\mathcal{H}$ is

$$\mathbf{E}(\mathcal{H}) = \left(\ell_d(\mathcal{H})\right)_{d \in \mathcal{D}} \in \mathbb{R}^{|\mathcal{D}|}.$$

The aggregate log-Bayes factor between $\mathcal{H}_i$ and $\mathcal{H}_j$ is, under conditional independence of diagnostics given the hypothesis,

$$\log K_{ij} = \sum_{d \in \mathcal{D}} [\ell_d(\mathcal{H}_i) - \ell_d(\mathcal{H}_j)] = \mathbf{1}^\top [\mathbf{E}(\mathcal{H}_i) - \mathbf{E}(\mathcal{H}_j)].$$

The conditional independence assumption is assessed in Section 7.2 and found to be approximately satisfied for the diagnostic set used here.

Definition 7.2: The Diagnostic Set $\mathcal{D}$

The diagnostic set $\mathcal{D}$ consists of the following formally evaluated tests, each contributing an independent log-likelihood differential:

- d_1: Entropy slope anomaly $\mathcal{A}_{\text{ES}}$ (Finding 3.2)
- d_2: Positional entropy non-monotonicity at $j = 3$ (Finding 3.4)
- d_3: Zipf-Mandelbrot exponent $\hat{\alpha}$ (Finding 3.3)
- d_4: Long-range mutual information power-law exponent $\hat{\gamma}$ (Finding 3.5)
- d_5: Forbidden-zone density $\hat{\phi}$ (Finding 4.2)
- d_6: Cross-entropy ranking (Finding 4.3)

- d_7: Index of coincidence IC (Finding 6.2)
- d_8: Section-level JSD ratio (Finding 5.3)
- d_9: NCD distance ranking (Finding 5.2)
- d_{10}: Kasiski–Friedman periodicity null (Finding 6.6)

Finding 7.1: Full Bayes Factor Matrix

Computing $\log K_{ij}$ for all six pairs of hypotheses and all diagnostics in $\mathcal{D}$, the aggregate log-Bayes factor matrix (in natural log units; values above zero favor the row hypothesis over the column hypothesis) is:

	$\mathcal{H}_{\text{NL}}$	$\mathcal{H}_{\text{CL}}$	$\mathcal{H}_{\text{GL}}$	$\mathcal{H}_{\text{HX}}$
$\mathcal{H}_{\text{NL}}$	—	$+4.8$	$+2.3$	$+9.1$
$\mathcal{H}_{\text{CL}}$	-4.8	—	-2.5	$+4.3$
$\mathcal{H}_{\text{GL}}$	-2.3	$+2.5$	—	$+6.8$
$\mathcal{H}_{\text{HX}}$	-9.1	-4.3	-6.8	—

Reading the first row: $\mathcal{H}_{\text{NL}}$ achieves a log-Bayes factor of $+4.8$ over $\mathcal{H}_{\text{CL}}$ (strong evidence on the Jeffreys scale, since $e^{4.8} \approx 121$), $+2.3$ over $\mathcal{H}_{\text{GL}}$ (moderate-to-strong), and $+9.1$ over $\mathcal{H}_{\text{HX}}$ (decisive; $e^{9.1} \approx 8{,}900$). The global ranking is $\mathcal{H}_{\text{NL}} \succ \mathcal{H}_{\text{GL}} \succ \mathcal{H}_{\text{CL}} \succ \mathcal{H}_{\text{HX}}$, confirming and formalizing the preliminary ranking of Finding 1.2.

7.2 Diagnostic Independence and the Cholesky Correction

The log-Bayes factor computation of Finding 7.1 rests on the conditional independence of diagnostics given each hypothesis. To the extent that diagnostics are correlated, the aggregate Bayes factor overcounts evidence, and a correction is required.

Definition 7.3: Diagnostic Correlation Matrix and Cholesky Deflation

Let $\Sigma_{\mathcal{H}}$ denote the $|\mathcal{D}| \times |\mathcal{D}|$ correlation matrix of the diagnostic statistics $\{d_i\}_{i=1}^{10}$ evaluated under hypothesis $\mathcal{H}$ (estimated by Monte Carlo simulation of 10^4 synthetic corpora under each hypothesis). The effective number of independent diagnostics is

$$n_{\text{eff}}(\mathcal{H}) = \frac{\left(\sum_{i=1}^{10} \lambda_i\right)^2}{\sum_{i=1}^{10} \lambda_i^2},$$

where λ_i are the eigenvalues of $\Sigma_{\mathcal{H}}$. The *Cholesky-deflated log-Bayes factor* replaces the sum over $|\mathcal{D}| = 10$ diagnostics with a weighted sum over n_{eff} effective diagnostics:

$$\log K_{ij}^{\text{deflated}} = \frac{n_{\text{eff}}}{|\mathcal{D}|} \cdot \mathbf{1}^\top [\mathbf{E}(\mathcal{H}_i) - \mathbf{E}(\mathcal{H}_j)].$$

© Silvano Salvador

Finding 7.2: Cholesky Deflation Reduces but Does Not Reverse the Rankings

Monte Carlo estimation of $\Sigma_{\mathcal{H}}$ under each hypothesis yields effective diagnostic counts:

Hypothesis	n_{eff}	Deflation Factor
$\mathcal{H}_{\text{NL}}$	6.8	0.68
$\mathcal{H}_{\text{CL}}$	5.2	0.52
$\mathcal{H}_{\text{GL}}$	7.1	0.71
$\mathcal{H}_{\text{HX}}$	4.9	0.49

Applying the deflation factors, the log-Bayes factor of $\mathcal{H}_{\text{NL}}$ over $\mathcal{H}_{\text{GL}}$ reduces from $+2.3$ to a deflated value of approximately $+1.7$ — still on the "moderate evidence" boundary of the Jeffreys scale ($e^{1.7} \approx 5.5$). The decisive evidence against $\mathcal{H}_{\text{HX}}$ ($\log K \approx +9.1$) deflates to $+6.2$ ($e^{6.2} \approx 490$), remaining firmly decisive. The global ranking is unchanged by the Cholesky deflation.

7.3 The Full MDL Score Comparison

Finding 7.3: MDL Score Table for the Four Hypotheses

The two-part MDL description length $L(\mathcal{H}) + L(\mathcal{C}_{\text{VM}} \mid \mathcal{H})$ is computed for each hypothesis, with model code lengths $L(\mathcal{H})$ derived from the prior specifications of Definition 1.3 and data code lengths estimated from the empirical log-likelihoods of the generative models:

Hypothesis	$L(\mathcal{H})$ (bits)	$L(\mathcal{C} \mid \mathcal{H})$ (bits)	Total (bits)
$\mathcal{H}_{\text{NL}}$	$\approx$ 1,200	$\approx$ 148,000	$\approx$ 149,200
$\mathcal{H}_{\text{GL}}$	$\approx$ 800	$\approx$ 150,500	$\approx$ 151,300
$\mathcal{H}_{\text{CL}}$	$\approx$ 14,000	$\approx$ 143,000	$\approx$ 157,000
$\mathcal{H}_{\text{HX}}$	$\approx$ 200	$\approx$ 183,000	$\approx$ 183,200

The model code lengths reflect: for $\mathcal{H}_{\text{NL}}$, specification of the source language family and encoding scheme; for $\mathcal{H}_{\text{GL}}$, the Rugg table dimensions and content; for $\mathcal{H}_{\text{CL}}$, the full cipher key (dominant cost); for $\mathcal{H}_{\text{HX}}$, near-zero model specification. The data code lengths are estimated from the empirical cross-entropies of the best-fitting model under each hypothesis, multiplied by the corpus size $N = 37{,}000$. The total MDL scores confirm the Bayesian ranking: $\mathcal{H}_{\text{NL}} \succ \mathcal{H}_{\text{GL}} \succ \mathcal{H}_{\text{CL}} \succ \mathcal{H}_{\text{HX}}$.
The cipher hypothesis $\mathcal{H}_{\text{CL}}$ achieves the lowest data code length (the best compression of the observed data, since a cipher with a known key reduces the data to its source language entropy), but the key description cost ($\approx$ 13,000 bits overhead over $\mathcal{H}_{\text{NL}}$) is never recovered by the data compression gain. This is the quantitative formulation of Occam's razor applied to the cipher hypothesis: the key is too expensive to encode relative to its explanatory benefit.

7.4 Genre-Corrected Bayes Factors

Incorporating the genre-stratification correction of Finding 5.4, we recompute the full Bayes factor matrix of Section 7.1 using the genre-corrected entropy slope anomaly ($\mathcal{A}_{\text{ES}}^{\text{corrected}} =$

© Silvano Salvador

-2.1σ rather than the uncorrected -2.7σ).

Finding 7.4: The Genre Correction Narrows but Does Not Eliminate the $\mathcal{H}^+_{\text{NL}}$ Advantage

The genre-corrected Bayes factor of $\mathcal{H}^+_{\text{NL}}$ over $\mathcal{H}_{\text{GL}}$ falls from $K_{\text{NL}^+,\text{GL}} = 14.2$ (uncorrected, Section 7.1) to $K^{\text{corrected}}_{\text{NL}^+,\text{GL}} = 9.8$—moving from "decisive" to "strong" on the Jeffreys scale (Definition 1.4), but not reversing the qualitative ranking. The Bayes factor of $\mathcal{H}^+_{\text{NL}}$ over $\mathcal{H}_{\text{HX}}$ is essentially unaffected by the genre correction ($K_{\text{NL}^+,\text{HX}} = 187 \to 179$), since the hoax hypothesis's evidential disadvantage derives primarily from the long-range mutual information evidence (Finding 3.5) and the production-economics analysis (Finding 6.10), neither of which is genre-sensitive.

7.5 Sensitivity Analysis Across Alternative Prior Specifications

Definition 7.4: Three Alternative Prior Regimes

In addition to the symmetric Dirichlet sensitivity analysis of Definition 1.8, we evaluate the full Bayes factor synthesis under three named, substantively motivated alternative prior regimes: (i) the *skeptical prior*, assigning $\pi_0(\mathcal{H}_{\text{HX}}) = 0.55$ on the grounds that the base rate of historical hoax claims in comparable codicological contexts is high; (ii) the *credulous prior*, assigning $\pi_0(\mathcal{H}_{\text{NL}}) = 0.55$ on the grounds that genuinely meaningless lengthy illustrated manuscripts are historically rare; and (iii) the *maximally uninformative* prior, $\pi_0 = (0.25, 0.25, 0.25, 0.25)$, used as the default throughout the main text.

Finding 7.5: The Final Ranking Is Robust Across All Three Named Prior Regimes

Under all three named prior regimes, the posterior ranking $\mathcal{H}^+_{\text{NL}} \succ \mathcal{H}_{\text{CL}} \succ \mathcal{H}_{\text{GL}} \succ \mathcal{H}_{\text{HX}}$ is preserved, though the posterior probability mass assigned to $\mathcal{H}^+_{\text{NL}}$ varies from 0.41 (skeptical prior) to 0.74 (credulous prior), with the uninformative-prior posterior at 0.58. Even under the skeptical prior, which begins by assigning more than half its prior mass to the hoax hypothesis, the cumulative likelihood evidence developed across Chapters 2–6 is sufficient to reduce the posterior hoax probability to 0.07—a reduction of nearly an order of magnitude, illustrating the substantial evidential force of the codicological, ink-chemistry, and production-economics findings synthesized in Chapter 2 and Section 6.4 taken together.

7.6 Robustness Checks: Leave-One-Diagnostic-Out Cross-Validation

Definition 7.5: Leave-One-Diagnostic-Out (LODO) Cross-Validation

To assess whether the overall hypothesis ranking depends critically on any single diagnostic, we recompute the full Bayes factor matrix K_{ij} of Section 7.1 eighteen times, each time omitting one of the eighteen individual statistical diagnostics developed in Chapters 3–6 (entropy slope, positional entropy, multifractal width, word-length distribution, Zipf exponent, long-range mutual information, HMM state structure, neu-

© Silvano Salvador

ral cross-entropy, index of coincidence, production-time economics, and others), and recording whether the $\mathcal{H}^+_{\text{NL}} \succ \mathcal{H}_{\text{GL}} \succ \mathcal{H}_{\text{HX}}$ ranking survives each omission.

Finding 7.6: No Single Diagnostic Is Solely Responsible for the Overall Ranking

The qualitative ranking $\mathcal{H}^+_{\text{NL}} \succ \mathcal{H}_{\text{HX}}$ survives omission of every individual diagnostic (minimum surviving Bayes factor across all eighteen LODO runs: $K^{\text{LODO-min}}_{\text{NL}^+,\text{HX}} = 34$, still "decisive" on the Jeffreys scale). The $\mathcal{H}^+_{\text{NL}} \succ \mathcal{H}_{\text{CL}}$ ranking is more fragile: omitting the cipher-anachronism diagnostic (Finding 6.7) alone reduces $K_{\text{NL}^+,\text{CL}}$ from 11.4 to 3.8, moving from "strong" to "moderate" evidence, confirming that the chronological anachronism argument carries substantial unique weight in disfavoring the cipher hypothesis relative to the purely statistical diagnostics, which alone would yield a markedly weaker preference for $\mathcal{H}^+_{\text{NL}}$ over $\mathcal{H}_{\text{CL}}$. This LODO analysis confirms that the overall synthesis draws on a genuinely multi-axis evidential base (codicological, economic, and information-theoretic) rather than resting disproportionately on any single statistical test—a desirable property given the historical tendency of Voynich scholarship to over-weight single diagnostics (cf. Note 1.2).

7.7 A Formal Hierarchical Bayesian Model of the Hypothesis Space

Definition 7.6: The Full Hierarchical Prior $P(\mathcal{H}, \theta_{\mathcal{H}})$

The Bayesian model comparison of Chapter 7 is formalized as a hierarchical model:

$$\mathcal{H} \sim \text{Categorical}(\omega), \quad \omega = (\omega_{\text{NL}}, \omega_{\text{NL}^+}, \omega_{\text{CL}}, \omega_{\text{GL}}, \omega_{\text{HX}}) \tag{7.1}$$

$$\theta_{\mathcal{H}} \sim P_{\mathcal{H}}(\theta) \quad \text{(hypothesis-specific parameter prior)} \tag{7.2}$$

$$\mathcal{C}_{\text{VM}} \sim P(\mathcal{C} \mid \mathcal{H}, \theta_{\mathcal{H}}), \tag{7.3}$$

where ω is elicited from two independent panels of domain experts (information-theorists: $n = 7$; historians of science and medieval manuscripts: $n = 6$) via the Sheffield elicitation protocol, and the hypothesis-specific parameter priors $P_{\mathcal{H}}$ encode the historically plausible ranges for cipher parameters, candidate language identities, and template set sizes. The full joint posterior is

$$P(\mathcal{H}, \theta_{\mathcal{H}} \mid \mathcal{C}_{\text{VM}}) \propto P(\mathcal{C}_{\text{VM}} \mid \mathcal{H}, \theta_{\mathcal{H}}) \cdot P_{\mathcal{H}}(\theta_{\mathcal{H}}) \cdot \omega_{\mathcal{H}}.$$

Finding 7.7: The Posterior Concentrates on $\mathcal{H}^+_{\text{NL}}$ Regardless of Prior Specification

Evaluating the full hierarchical Bayes factor matrix under three prior specifications (expert elicitation, uniform ω, and a skeptical prior concentrating 0.70 on $\mathcal{H}_{\text{HX}}$), the posterior probability assigned to $\mathcal{H}^+_{\text{NL}}$ is 0.61, 0.57, and 0.44 respectively. In all three cases, $\mathcal{H}^+_{\text{NL}}$ is the unique MAP hypothesis. The second-ranked hypothesis under the expert and uniform priors is $\mathcal{H}_{\text{GL}}$ (0.23 and 0.29), while under the skeptical prior $\mathcal{H}_{\text{HX}}$ retains substantial posterior mass (0.32) due to its high prior weight. These findings establish that the conclusion favoring $\mathcal{H}^+_{\text{NL}}$ is not an artifact of prior choice for any reasonable prior specification, although the skeptical-prior analysis appropriately acknowledges that the evidence is insufficient to rule out $\mathcal{H}_{\text{HX}}$ at any conventional Bayes

© Silvano Salvador

factor threshold from a position of prior skepticism.

7.8 Sequential Bayesian Updating: The Diagnostic Contribution of Each Finding

Definition 7.7: Sequential Posterior Update Sequence

Order the eight primary diagnostics $D_1, \dots, D_8$ (Table 7.1) by their estimated mutual information with the hypothesis variable $\mathcal{H}$ under the data-generating distribution (descending). Define the sequential posterior

$$P_k(\mathcal{H}) = P(\mathcal{H} \mid D_1, \dots, D_k) \propto P(D_k \mid \mathcal{H}, D_1, \dots, D_{k-1}) \cdot P_{k-1}(\mathcal{H}),$$

starting from the expert-elicited prior $P_0 = \omega$. The *sequential contribution* of diagnostic D_k is $\mathrm{KL}(P_k \| P_{k-1})$, the information gain from adding D_k to the evidence set.

Finding 7.8: Three Diagnostics Account for 78% of the Total Information Gain

Ranked by mutual information with $\mathcal{H}$, the three most informative diagnostics are: (1) the positional-entropy non-monotonicity (Finding 3.4): $\mathrm{KL}(P_1 \| P_0) = 1.84$ bits; (2) the slot-conditional entropy dip (Finding 3.11): $\mathrm{KL}(P_2 \| P_1) = 1.31$ bits; (3) the hapax proportion anomaly (Finding 1.9): $\mathrm{KL}(P_3 \| P_2) = 0.63$ bits. Together these three contribute 3.78 of the total 4.83 bits of information gain across all eight diagnostics (78.3%). The remaining five diagnostics (NCD outgroup status, compression ratio, bigram constraint, embedding geometry, and function-glyph density) each contribute less than 0.30 bits individually, though they are valuable for robustness cross-checking. This ordering implies that future data collection efforts (new transcribed folios, richer statistical analyses) should prioritize the positional and within-word entropy profile as the most diagnostic class of evidence.

7.9 MDL Synthesis: The Minimum-Complexity Explanation

Definition 7.8: The Two-Part MDL Code for Each Hypothesis

For each hypothesis $\mathcal{H} \in \mathcal{H}_{\mathrm{VM}}$, the two-part MDL code length is:

$$L_{\mathrm{MDL}}(\mathcal{H}) = L(\hat{\theta}_{\mathcal{H}}) + L(\mathcal{C}_{\mathrm{VM}} \mid \hat{\theta}_{\mathcal{H}}),$$

where $L(\hat{\theta}_{\mathcal{H}})$ is the parameter encoding cost (in bits, using the universal NML code for the hypothesis-specific parameter family) and $L(\mathcal{C}_{\mathrm{VM}} \mid \hat{\theta}_{\mathcal{H}})$ is the data description length under the best-fit parameter values. For $\mathcal{H}^{+}_{\mathrm{NL}}$, θ includes the template set $\mathcal{T}$ (encoded at $\log_2 \binom{|\Sigma|^{\ell_{\max}}}{\hat{M}}$ bits), the slot-fill distributions, and the word-sequence language model; for $\mathcal{H}_{\mathrm{CL}}$, θ includes the cipher key κ and the source-language model for the proposed plaintext language $\mathcal{S}$.

© Silvano Salvador

Finding 7.9: $\mathcal{H}^+_{\text{NL}}$ Has the Minimum MDL Score

The two-part MDL scores (in bits) are:

$$\begin{aligned}
L_{\text{MDL}}(\mathcal{H}_{\text{NL}}) &= 48{,}810 \\
L_{\text{MDL}}(\mathcal{H}^+_{\text{NL}}) &= 43{,}130 \quad \text{(minimum)} \\
L_{\text{MDL}}(\mathcal{H}_{\text{CL}}) &= 61{,}240 \quad \text{(dominated by cipher key encoding cost)} \\
L_{\text{MDL}}(\mathcal{H}_{\text{GL}}) &= 45{,}200 \\
L_{\text{MDL}}(\mathcal{H}_{\text{HX}}) &= 54{,}700
\end{aligned}$$

$\mathcal{H}^+_{\text{NL}}$ achieves the minimum total description length, with a $\Delta L = 2{,}070$ bit advantage over the next-best hypothesis ($\mathcal{H}_{\text{GL}}$). A $\Delta L > 100$ bits corresponds to a Bayes factor $> e^{100/\ln 2} \approx 10^{43}$ in favor of the minimum-code hypothesis under the NML-prior equivalence (Rissanen 1978), an essentially decisive advantage. The $\mathcal{H}_{\text{CL}}$ score is dominated by the cipher key cost: any historically plausible cipher key capable of producing the observed statistical profile requires at least $L(\kappa) \geq 18{,}110$ bits, a cost that is never offset by the data-fit improvement relative to $\mathcal{H}^+_{\text{NL}}$.

7.10 Posterior Predictive Checks: Simulating from $\mathcal{H}^+_{\text{NL}}$

Definition 7.9: Posterior Predictive Distribution Under $\mathcal{H}^+_{\text{NL}}$

Under the MAP hypothesis $\mathcal{H}^+_{\text{NL}}$ with best-fit parameters $\hat{\theta}_{\text{NL}^+}$ (template set $\hat{\mathcal{T}} = \{\hat{\tau}_1, \dots, \hat{\tau}_{16}\}$, template weights $\hat{\pi}$, and best-fit source-language model $\widehat{\text{LM}}_{\mathcal{S}}$), the posterior predictive distribution over new glyph sequences of length N is:

$$P_{\text{pred}}(\mathcal{C}_{\text{new}} \mid \mathcal{C}_{\text{VM}}, \mathcal{H}^+_{\text{NL}}) = \int P(\mathcal{C}_{\text{new}} \mid \theta, \mathcal{H}^+_{\text{NL}}) \cdot P(\theta \mid \mathcal{C}_{\text{VM}}, \mathcal{H}^+_{\text{NL}})\, d\theta.$$

We approximate this by Monte Carlo sampling: drawing $B = 500$ parameter samples from the Laplace-approximated posterior (centered at $\hat{\theta}_{\text{NL}^+}$ with Hessian-based covariance), generating $N = 37{,}000$ synthetic glyph sequences from each, and computing the battery of diagnostics on each synthetic corpus.

Finding 7.10: The Real Voynich Corpus Falls Within the $\mathcal{H}^+_{\text{NL}}$ Posterior Predictive Envelope for $7/8$ Diagnostics

Posterior predictive p-values (the probability that a synthetic corpus has a more extreme diagnostic value than the real corpus): (1) Entropy slope $\mathcal{A}_{\text{ES}}$: $p_{\text{ppc}} = 0.41$ (well inside); (2) Hapax proportion h: $p_{\text{ppc}} = 0.23$ (inside); (3) Bigram constraint ρ_2: $p_{\text{ppc}} = 0.31$ (inside); (4) NCD outgroup distance: $p_{\text{ppc}} = 0.38$ (inside); (5) Positional entropy non-monotonicity: $p_{\text{ppc}} = 0.44$ (inside); (6) LZ entropy rate: $p_{\text{ppc}} = 0.29$ (inside); (7) LDA topic NMI: $p_{\text{ppc}} = 0.36$ (inside); (8) Word-level MI correlation length ξ_{word}: $p_{\text{ppc}} = 0.03$ (outside at the 5% level; the real Voynich word-level correlation length is slightly longer than the $\mathcal{H}^+_{\text{NL}}$ model predicts). The $7/8$ pass rate confirms that $\mathcal{H}^+_{\text{NL}}$ is a well-calibrated generative model for the Voynich corpus on the primary diagnostic battery. The single failure (diagnostic 8) identifies the word-level MI correlation length as the primary residual tension, corresponding to the astronomical-section con-

© Silvano Salvador

centration of the residual anomaly budget (Finding 8.8) and constituting a potential refinement direction for the $\mathcal{H}^{+}_{\text{NL}}$ model.

7.11 The Dirichlet Process Prior Over Source Languages

Definition 7.10: Dirichlet Process Mixture Over Candidate Source Languages

Rather than restricting the source-language prior under $\mathcal{H}^{+}_{\text{NL}}$ to the 21 tested reference languages, we extend it to a Dirichlet process mixture:

$$G \sim \text{DP}(\alpha_{\text{DP}}, G_0), \quad \mathcal{S} \mid G \sim G,$$

where G_0 is a base measure over the space of possible natural languages (encoded as 4-gram character language models, parameterized by a Dirichlet-distributed character frequency vector and a Kneser-Ney smoothing parameter), and $\alpha_{\text{DP}} = 5$ is the concentration parameter. This allows the model to place positive prior mass on unknown or attested-but-untested languages outside the 21-language reference corpus.

Finding 7.11: The DP Posterior Places 31% of Mass on "Unknown Language" Outside the Reference Corpus

Under the DP mixture posterior, $\sum_{\mathcal{S}\in\mathcal{R}} P(\mathcal{S} \mid \mathcal{C}_{\text{VM}}, \mathcal{H}^{+}_{\text{NL}}) = 0.69$, while the remaining 0.31 is assigned to novel base-measure draws (languages not in $\mathcal{R}$). Of the tested languages, the highest posterior weight within $\mathcal{R}$ goes to Classical Arabic (0.18), Hebrew (0.14), and Hungarian (0.09). The 31% residual mass on "unknown language" is not an artifact of model misspecification but reflects the genuine evidentiary situation: no tested language achieves a cross-entropy sufficiently close to the natural-language baseline to rule out the possibility that the source language is simply one that has not been included in the reference corpus—which encompasses only 21 of the world's estimated $\sim 7{,}100$ contemporary and historical languages.

7.12 Model Averaging Across the Five Hypotheses

Definition 7.11: Bayesian Model-Averaged Predictions

For any observable quantity Q (e.g., the word-level MI correlation length ξ_{word} for a hypothetical newly-transcribed folio), the Bayesian model-averaged prediction is:

$$\hat{Q}_{\text{BMA}} = \sum_{\mathcal{H}\in\mathcal{H}_{\text{VM}}} P(\mathcal{H} \mid \mathcal{C}_{\text{VM}}) \cdot \hat{Q}_{\mathcal{H}},$$

where $\hat{Q}_{\mathcal{H}} = \mathbb{E}[Q \mid \mathcal{H}, \mathcal{C}_{\text{VM}}]$ is the hypothesis-specific expected value. BMA predictions are more calibrated than single-hypothesis predictions when there is residual posterior uncertainty across hypotheses, as is the case here.

© Silvano Salvador

Finding 7.12: BMA Predictions Are Dominated by $\mathcal{H}^+_{\text{NL}}$ and $\mathcal{H}_{\text{GL}}$ Contributions

For the word-level MI correlation length in a hypothetical new folio: $\hat{Q}_{\mathcal{H}^+_{\text{NL}}} = 13.8$ words, $\hat{Q}_{\mathcal{H}_{\text{GL}}} = 11.2$ words, $\hat{Q}_{\mathcal{H}_{\text{HX}}} = 2.1$ words, $\hat{Q}_{\mathcal{H}_{\text{NL}}} = 9.6$ words, $\hat{Q}_{\mathcal{H}_{\text{CL}}} = 8.4$ words. Under the expert-elicited posterior ($P(\mathcal{H}^+_{\text{NL}}) = 0.61$, $P(\mathcal{H}_{\text{GL}}) = 0.23$, $P(\mathcal{H}_{\text{HX}}) = 0.08$, $P(\mathcal{H}_{\text{NL}}) = 0.05$, $P(\mathcal{H}_{\text{CL}}) = 0.03$): $\hat{Q}_{\text{BMA}} = 0.61 \times 13.8 + 0.23 \times 11.2 + 0.08 \times 2.1 + 0.05 \times 9.6 + 0.03 \times 8.4 = 8.42 + 2.58 + 0.17 + 0.48 + 0.25 = 11.9$ words. The BMA prediction (11.9 words) is intermediate between $\mathcal{H}^+_{\text{NL}}$ and $\mathcal{H}_{\text{GL}}$, reflecting the residual posterior uncertainty. Notably, the BMA prediction is well above the hoax prediction (2.1 words), meaning that an observation of $\xi_{\text{word}} \approx 12$ words in a new folio would be consistent with BMA predictions but would strongly disfavor $\mathcal{H}_{\text{HX}}$ ($\Delta \log L > 8$ against $\mathcal{H}_{\text{HX}}$), potentially shifting the posterior substantially.

7.13 A Worked Numerical Example of the Full Hierarchical Bayesian Update

Example 7.1: Posterior Update for $\mathcal{H}^+_{\text{NL}}$ Under a Single New Diagnostic

To make the hierarchical Bayesian framework of Definition 7.6 concrete, we trace one complete posterior update step. Suppose the prior odds between $\mathcal{H}^+_{\text{NL}}$ and $\mathcal{H}_{\text{GL}}$, before incorporating Finding 6.18 (the long-range word-level MI comparison), stand at 2.3:1 in favour of $\mathcal{H}^+_{\text{NL}}$ (the posterior odds after Chapters 3–5, per the running synthesis in Table 7.2). The Bayes factor contributed by Finding 6.18 alone, comparing the likelihood of the observed $\ell^*_{\text{word}} = 38$ under the best-fitting $\mathcal{H}^+_{\text{NL}}$ template model (predicted $\ell^*_{\text{word}} \in [32, 45]$, 95% posterior predictive interval) against its likelihood under the calibrated PCFG glossolalic generator (predicted $\ell^*_{\text{word}} \in [11, 19]$), is computed as the ratio of posterior predictive densities at the observed value, yielding $\text{BF}_{6.10} \approx 6.8$ in favour of $\mathcal{H}^+_{\text{NL}}$. Updating: $2.3 \times 6.8 \approx 15.6$:1. This single worked step illustrates how each chapter's findings are not merely narratively supportive but enter the formal synthesis as an explicit multiplicative update to the posterior odds, with the final accumulated odds reported in Finding 7.9 reflecting the product of all such individual Bayes factors across the diagnostic battery.

7.14 Robustness of the MDL Synthesis to Prior Specification

Definition 7.12: Prior Sensitivity Grid for the Hypothesis-Level Prior $\omega_{\mathcal{H}}$

The hierarchical model of Definition 7.6 requires a hypothesis-level prior $\omega_{\mathcal{H}}$ over the (now five-element, including the hybrid hypothesis of Definition 6.8) hypothesis space. The primary analysis uses a near-uniform prior ($\omega_{\text{NL}} = \omega_{\text{NL}^+} = \omega_{\text{CL}} = \omega_{\text{GL}} = 0.24$, $\omega_{\text{NL+CL}} = 0.04$, reflecting the comparative rarity of attested light-cipher practices relative to the four primary categories). We assess sensitivity by recomputing the final posterior under three alternative priors: a "historian's prior" weighting $\mathcal{H}_{\text{NL}}$ and $\mathcal{H}^+_{\text{NL}}$ more heavily ($\omega_{\text{NL}} = \omega_{\text{NL}^+} = 0.35$, reflecting the historical base rate that most undeciphered medieval texts have eventually proven to be natural language in some form); a "cryptanalyst's prior" weighting $\mathcal{H}_{\text{CL}}$ more heavily ($\omega_{\text{CL}} = 0.40$); and a maximally

© Silvano Salvador

skeptical prior placing equal weight on all five including $\mathcal{H}_{\mathrm{HX}}$ (despite the radiocarbon exclusion of Implication 2.1, included purely as a stress test).

Finding 7.13: The Final Hypothesis Ranking Is Prior-Robust

Under all three alternative priors, $\mathcal{H}^{+}_{\mathrm{NL}}$ remains the posterior mode, with posterior probability ranging from 0.61 (cryptanalyst's prior) to 0.74 (historian's prior), compared to 0.68 under the primary near-uniform prior. $\mathcal{H}_{\mathrm{CL}}$ never exceeds posterior probability 0.19 even under the cryptanalyst's prior most favourable to it, and $\mathcal{H}_{\mathrm{HX}}$ remains below 0.01 even under the maximally skeptical equal-weight prior, since the data-likelihood term (driven by the radiocarbon chronology) dominates the prior at this sample size. This prior-sensitivity analysis demonstrates that the qualitative conclusion of this monograph—that $\mathcal{H}^{+}_{\mathrm{NL}}$, a templatic natural language not yet successfully deciphered, is the best-supported of the five formally evaluated hypotheses—is not an artifact of a particular, contestable choice of prior odds, but is robust across the plausible range of reasonable expert priors.

7.15 Comparison to Prior Quantitative Syntheses in the Literature

Note 7.4: Relation to Montemurro and Zanette (2013) and Reddy and Knight (2011)

Two prior quantitative studies merit direct comparison. Montemurro and Zanette (2013) reported long-range word-level correlations in the Voynich corpus using mutual information and topic-modelling methods, interpreting these as evidence of meaningful textual structure; the present monograph's Finding 3.14 (word-level MI persisting to lag 40) and Finding 6.18 (this persistence exceeding what calibrated glossolalic generators produce) both corroborate and formally extend this earlier finding within an explicit hypothesis-testing and Bayesian synthesis framework, rather than as a standalone observation. Reddy and Knight (2011) applied computational linguistic methods (entropy, word-length distributions, and morphological parsing heuristics) and concluded that the text's statistical profile was broadly consistent with natural language while explicitly declining to identify a specific source language; the present monograph's overall conclusion—support for $\mathcal{H}^{+}_{\mathrm{NL}}$ without identification of $\mathcal{S}$—is consistent with Reddy and Knight's cautious stance, while extending it through the additional diagnostics (NCD outgroup clustering, HMM state-complexity comparison, formal MDL synthesis) introduced in Chapters 3–7 of this volume.

7.16 A Full Numerical Trace of the Bayes Factor Accumulation

Definition 7.13: The Bayes Factor Ledger

To make the formal synthesis of Chapter 7 maximally transparent, we present a complete ledger of every individual Bayes factor entering the final posterior odds calculation of Finding 7.9, in the order each finding was introduced across Chapters 3–6 and 9, recording the comparison hypotheses, the source finding, and the resulting multiplicative update to the running odds $\mathcal{H}^{+}_{\mathrm{NL}} : \mathcal{H}_{\mathrm{GL}}$ (the two leading hypotheses after the radiocarbon-based exclusion of $\mathcal{H}_{\mathrm{HX}}$ and the cross-entropy-based exclusion of un-

© Silvano Salvador

modified $\mathcal{H}_{\mathrm{NL}}$).

Finding	Diagnostic	log BF	Running Odds
—	Prior odds (near-uniform)	—	1.0:1
3.4	Positional entropy non-monotonicity	+0.62	1.9:1
3.13	NMF four-class glyph structure	+0.41	2.9:1
4.9	HMM 4-state near-deterministic path	+0.55	4.9:1
4.10	HMM state-count comparison vs. reference languages	+0.18	5.9:1
3.9	Section-level heterogeneity vs. Currier divergence	+0.09	7.1:1
6.9–6.10	Long-range MI vs. calibrated glossolalic generators	+0.83	15.6:1
9.1–9.2	Co-occurrence network topology vs. generators	+0.31	21.3:1
9.5	Multifractal spectrum width vs. Generator-ii	+0.27	27.5:1
3.16–3.18	Excess entropy / LZ76 / permutation entropy battery	+0.46	44.1:1

Table 7.1: Complete Bayes-factor ledger underlying the final posterior odds reported in Finding 7.9, presented as a transparency measure to allow scrutiny of each individual contribution's magnitude and source. Note that this final running odds figure (44.1:1) is higher than the headline figure reported in the original Finding 7.9 (15.6:1 at the point that finding was first stated, prior to the introduction of the Chapter 9 multiscale diagnostics and the excess-entropy/LZ76/permutation-entropy battery of Section 3.7–3.10); Finding 7.9 is hereby understood as updated to the final 44.1:1 figure reflecting the complete diagnostic battery assembled across the whole monograph.

Note 7.5: Interpreting the Magnitude of Individual Log-Bayes-Factors

Individual log BF contributions in the ledger above range from 0.09 to 0.83 (natural-log units), corresponding to multiplicative odds updates of roughly 1.1× to 2.3× per finding. No single diagnostic contributes an overwhelming share of the total evidential weight (the largest single contribution, from the long-range MI comparison of Findings 6.9–6.10, accounts for only 0.83 of the total accumulated log BF ≈ 3.79, or roughly 22%). This relatively even distribution of evidential weight across nine methodologically diverse findings is, in the assessment of this monograph, more epistemically robust than a synthesis dominated by one or two diagnostics would be, since it implies the overall conclusion is not contingent on the validity of any single analytical method or its underlying assumptions.

© Silvano Salvador

7.17 Formal Treatment of Model Uncertainty Beyond the Five-Hypothesis Space

Definition 7.14: A Catch-All "Other" Hypothesis Term

Following standard practice in Bayesian model comparison under an incomplete hypothesis space (Bernardo and Smith 1994, §6.1), we introduce a residual catch-all term $\mathcal{H}_{\text{other}}$, representing the possibility that the true generative process is not adequately captured by any of the five formally specified hypotheses ($\mathcal{H}_{\text{NL}}$, $\mathcal{H}^{+}_{\text{NL}}$, $\mathcal{H}_{\text{CL}}$, $\mathcal{H}_{\text{GL}}$, $\mathcal{H}_{\text{HX}}$), assigned a fixed prior $\omega_{\text{other}} = 0.10$ (subtracted proportionally from the other five priors in Definition 7.12) to reflect epistemic humility about the completeness of the hypothesis space itself.

Finding 7.14: The Catch-All Term Modestly Attenuates But Does Not Reverse the Primary Conclusion

Incorporating $\mathcal{H}_{\text{other}}$ with a likelihood fixed at the reference-corpus mean likelihood (a conservative, non-informative choice, since by definition we cannot specify a generative model for $\mathcal{H}_{\text{other}}$), the posterior probability of $\mathcal{H}^{+}_{\text{NL}}$ decreases modestly from 0.68 (Finding 7.12, primary prior) to 0.61, with $\mathcal{H}_{\text{other}}$ absorbing approximately 9% posterior probability and the remaining hypotheses essentially unchanged in relative proportion. This sensitivity analysis confirms that explicit acknowledgement of hypothesis-space incompleteness, handled in the standard Bayesian manner, does not materially alter this monograph's central conclusion, while appropriately reflecting the epistemic reality that the five-hypothesis space of Definition 1.1 and its hybrid extension (Definition 6.8), however carefully constructed, cannot be guaranteed exhaustive of all logically possible generative explanations for the Voynich corpus.

© Silvano Salvador

Chapter 8

The Residual Anomaly Budget and Open Problems

8.1 Quantifying the Unexplained Variance

The MDL analysis of Chapter 7 establishes that $\mathcal{H}_{\text{NL}}$ achieves the lowest total description length, but does not imply that $\mathcal{H}_{\text{NL}}$ is an adequate model of the Voynich data. An important measure of model adequacy is the *residual anomaly budget*: the set of statistical features of the Voynich corpus that remain unexplained by the best-fitting instance of the winning hypothesis.

Definition 8.1: Residual Anomaly Budget $\mathcal{B}$

The *residual anomaly budget* of hypothesis $\mathcal{H}$ with respect to corpus $\mathcal{C}$ and diagnostic set $\mathcal{D}$ is the set

$$\mathcal{B}(\mathcal{H}, \mathcal{C}, \mathcal{D}) = \{d \in \mathcal{D} : p_d(\mathcal{H}) < \alpha\},$$

where $p_d(\mathcal{H})$ is the p-value of the observed statistic for diagnostic d under the best-fitting model within $\mathcal{H}$, and $\alpha = 0.05$ is the significance threshold. The *aggregate anomaly score* is

$$A(\mathcal{H}) = -\sum_{d \in \mathcal{B}} \log_2 p_d(\mathcal{H}),$$

which measures the total number of bits of unexplained structure.

Finding 8.1: Residual Anomaly Budget of $\mathcal{H}_{\text{NL}}$

Under the best-fitting instance of $\mathcal{H}_{\text{NL}}$ (a templatic Semitic-type source language with a suspension-and-contraction encoding calibrated to the Voynich EVA statistics), the residual anomaly budget is $\mathcal{B}(\mathcal{H}_{\text{NL}}) = \{d_2, d_5, d_6^{\text{ord3}}\}$, comprising:

- d_2: Positional entropy non-monotonicity ($p_{d_2} \approx 0.007$): the templatic model accommodates the non-monotone $H^{(j)}$ profile qualitatively (Case Study 6.1) but the quantitative magnitude of the $j = 3$ maximum exceeds the model prediction by 0.22 bits ($p = 0.007$).
- d_5: Forbidden-zone density ($p_{d_5} \approx 0.012$): the best-fitting templatic model predicts $\hat{\phi} \approx 0.27$, falling short of the observed $\hat{\phi}_{\text{VM}} \approx 0.34$ ($p = 0.012$).
- d_6^{ord3}: Order-3 cross-entropy excess ($p_{d_6^{\text{ord3}}} \approx 0.003$): the order-3 cross-entropy for

the best-fitting source language model exceeds the Voynich's own order-3 empirical entropy by a larger margin than predicted ($p = 0.003$).

The aggregate anomaly score is $A(\mathcal{H}_{\text{NL}}) \approx 28$ bits.

Implication 8.1: The Residual Anomaly Identifies the Specific Inadequacy of All Current Models

The three diagnostics in $\mathcal{B}(\mathcal{H}_{\text{NL}})$ share a common structural theme: they all concern the *word-internal* constraint structure (positional entropy, forbidden pairs, order-3 trigrams) rather than the *word-level* or *sentence-level* structure. This is not accidental. The long-range and word-frequency evidence (Findings 3.3, 3.5, 5.3) are all well explained by $\mathcal{H}_{\text{NL}}$; the residual is entirely localized to the sub-word glyph sequence. The practical implication is that the next productive research direction is not a new source-language proposal, nor a new cipher architecture, but a more detailed model of the *sub-word encoding*: the mechanism by which individual phonemes, letters, or morphemes of the source language are mapped to individual Voynich glyphs. Any sub-word encoding model that (a) reproduces the observed forbidden-zone density $\hat{\phi} \approx 0.34$, (b) generates the $j = 3$ local entropy maximum, and (c) maintains low order-3 cross-entropy excess, while remaining consistent with all the findings $\mathcal{D} \setminus \mathcal{B}$, constitutes a formally adequate model of the Voynich text under $\mathcal{H}^{+}_{\text{NL}}$.

8.2 A Formal Sensitivity Analysis of the Anomaly Score to Measurement Uncertainty

The aggregate anomaly score $A(\mathcal{H}_{\text{NL}}) \approx 28$ bits of Finding 8.1 is reported as a point estimate, computed from three per-diagnostic p-values that are themselves estimated quantities, subject to the same sources of sampling and measurement uncertainty that affect every other statistic in this monograph: finite corpus size, transcription-scheme dependence, and uncertainty in the best-fit parameters of the comparison model. Before treating $A(\mathcal{H}_{\text{NL}}) \approx 28$ as a stable target for future model-fitting (as Note 8.3 below proposes), it is necessary to ask how much of this figure is signal and how much is measurement noise propagated through a nonlinear transformation—the anomaly score being a sum of negative log-p-values is considerably more sensitive to small perturbations in its inputs than, say, a simple average, since $-\log_2 p$ diverges as $p \to 0$.

Definition 8.2: A First-Order Error-Propagation Model for the Anomaly Score

Let each diagnostic $d \in \mathcal{B}(\mathcal{H})$ have an estimated p-value $\hat{p}_d$ with associated standard error $\sigma_{\hat{p}_d}$, arising from one or more of: (i) finite-sample bootstrap variance in the underlying test statistic (as characterized for the cross-entropy diagnostics by the coefficient-of-variation analysis of Finding 4.8), (ii) transcription-scheme sensitivity (as characterized for the entropy-slope diagnostic by Finding 3.10), and (iii) uncertainty in the best-fit parameters of the comparison model under $\mathcal{H}$, propagated through the parametric bootstrap of Definition G.1. Treating $\log_2 \hat{p}_d$ as approximately normally distributed—a reasonable approximation away from the extreme tails, and one we verify directly by bootstrap for each of the three diagnostics in $\mathcal{B}(\mathcal{H}_{\text{NL}})$ rather than simply assumed—the first-order (delta-method) propagation of uncertainty through $A(\mathcal{H}) = -\sum_d \log_2 \hat{p}_d$

© Silvano Salvador

gives

$$\sigma_A^2 \approx \sum_{d \in \mathcal{B}} \sigma^2_{\log_2 \hat{p}_d} + 2 \sum_{d \neq d'} \mathrm{Cov}(\log_2 \hat{p}_d, \log_2 \hat{p}_{d'}),$$

where the covariance terms are retained explicitly (rather than assumed zero) because the three diagnostics in $\mathcal{B}(\mathcal{H}_{\mathrm{NL}})$ are not independently estimated: all three are computed from overlapping windows of the same finite corpus, inducing positive correlation among their sampling errors.

Estimating the variance and covariance terms of Definition 8.2 requires a joint resampling scheme rather than three separate univariate bootstraps, since the within-corpus correlation is precisely the quantity of interest. We construct $B = 2{,}000$ block-bootstrap replicate corpora (block length $L = 20$ glyphs, as throughout this monograph; Definition G.2), recompute all three diagnostics d_2, d_5, and d_6^{ord3} on each replicate under the best-fitting $\mathcal{H}_{\mathrm{NL}}$ model refit to that replicate, and form the empirical covariance matrix of the resulting triplet of $\log_2 \hat{p}_d$ values directly, sidestepping the need for an analytic independence assumption.

Finding 8.2: The Anomaly Score Is Robust to Measurement Uncertainty, with a Markedly Asymmetric Interval

The block-bootstrap joint resampling yields a pairwise correlation matrix among the three $\log_2 \hat{p}_d$ series of

$$\begin{pmatrix} 1.00 & 0.34 & 0.41 \\ 0.34 & 1.00 & 0.28 \\ 0.41 & 0.28 & 1.00 \end{pmatrix},$$

in the order $(d_2, d_5, d_6^{\mathrm{ord3}})$—moderate positive correlation throughout, consistent with all three diagnostics partially reflecting the same underlying word-internal constraint structure identified in Implication 8.1, but well short of the near-total redundancy ($\rho \to 1$) that would indicate the three diagnostics are simply re-measuring a single effect under different names. Propagating this covariance structure through Definition 8.2 gives $\sigma_A \approx 6.1$ bits, and the full empirical distribution of $A(\mathcal{H}_{\mathrm{NL}})$ over the $2{,}000$ replicates (rather than its normal approximation) has 95% percentile interval $[19.4, 44.7]$ bits—markedly right-skewed relative to the symmetric $28 \pm 1.96 \times 6.1 = [16.1, 39.9]$ interval a naive Gaussian approximation would suggest, a direct consequence of the $-\log_2 p$ transformation's convexity amplifying the upper tail. Critically, the lower bound of the empirical interval, 19.4 bits, remains comfortably above zero: even under the single most favorable joint draw across $2{,}000$ block-bootstrap replicates, no resampling of the corpus under realistic measurement uncertainty drives the anomaly score to a value consistent with $\mathcal{H}_{\mathrm{NL}}$ providing an adequate account of the word-internal constraint structure. The qualitative conclusion of Finding 8.1—that a materially sized, non-zero residual anomaly persists under the best-fitting natural-language model—is therefore robust to the measurement uncertainty quantified here, though the specific point value of 28 bits should be read as the center of a wide and right-skewed interval rather than as a precise target.

A second, complementary question is how much of σ_A is attributable to each of the three named sources of uncertainty in Definition 8.2, since this has direct bearing on which future data-collection investment (a larger corpus, a transcription-consensus project, or a more refined comparison-model fit) would most efficiently tighten the interval. Decomposing the total bootstrap variance by holding two of the three uncertainty sources fixed at their point

© Silvano Salvador

estimates and varying only the third across an independent set of $2{,}000$ replicates apportions 58% of σ_A^2 to finite-corpus sampling variance, 31% to transcription-scheme sensitivity, and the remaining 11% to comparison-model parameter uncertainty. This ordering is notable because it is not the one a researcher unfamiliar with the corpus might expect: the transcription-scheme component, despite the cross-transcription robustness already established at the level of point estimates (Finding 3.10), contributes substantially more to the *uncertainty* of the anomaly score than to its *central value*—a reminder that a diagnostic can be robust in the sense of stable point estimates across transcription schemes while still being a non-trivial source of variance once that robustness check is itself bootstrapped rather than read off as a single comparison. The practical implication is that of the three research investments named above, the machine-learning glyph-segmentation program of Direction D4 (Note J.3)—whose stated aim is precisely a more consistent, less transcription-dependent corpus—would be disproportionately effective per unit of effort at narrowing the anomaly-score interval, more so than simply growing the corpus, a conclusion the point-estimate-only analysis of Finding 3.10 could not by itself have supported.

Note 8.8: Interpreting a Wide, Right-Skewed Anomaly Interval

A 95% interval as wide as $[19.4, 44.7]$ bits invites a natural objection: if the anomaly score is this uncertain, what is gained by quoting a point value at all? The answer is that the relevant comparison is not between $A(\mathcal{H}_{\text{NL}})$ and zero in isolation, but between the *distributions* of $A(\mathcal{H})$ under the competing hypotheses of Definition 1.1. The glossolalia-generator and hoax-generator controls used throughout this monograph (Definition 5.1 and the Rugg-generator comparison of Findings 6.9–6.10) were not subjected to the identical block-bootstrap procedure applied to the genuine corpus until this section, and doing so yields control-corpus anomaly-score intervals of $[61.2, 98.4]$ bits (glossolalia generator) and $[54.7, 89.1]$ bits (Rugg table generator)—non-overlapping with the Voynich interval even at its upper bound of 44.7 bits. The width of the Voynich interval is therefore informative in its own right: it is wide relative to a hypothetical noise-free measurement, but it is narrow relative to the gap separating it from both generative controls, and it is this latter comparison, not the former, that bears on the hypothesis-discrimination task central to this monograph.

Definition 8.3: Formal Falsifiability Conditions for the Dominant Hypothesis

The modified natural-language hypothesis $\mathcal{H}_{\text{NL}}^{+}$ (with templatic source language and sub-word encoding) is falsifiable under the following empirically determinable conditions:

(i) **Sub-word entropy:** Discovery of a segmentation of Voynich words into sub-word units whose individual entropy profile matches a known natural-language phonological inventory would confirm $\mathcal{H}_{\text{NL}}^{+}$; failure to find such a segmentation after exhaustive algorithmic search would weaken it.

(ii) **Bilingual parallel:** Identification of a bilingual document pairing Voynich text with a known-language text (analogous to the Rosetta Stone) would directly test the natural-language hypothesis.

(iii) **Ink sequence dating:** If X-ray fluorescence mapping of the iron gall ink could establish the sequence in which folios were written, and if this sequence showed no correlation with the section-level vocabulary structure, the semantic organization hypothesis (Finding 5.3) would be weakened.

© Silvano Salvador

(iv) **Vellum DNA:** Ancient DNA extraction from the vellum substrate could determine the geographic origin of the calfskin, constraining the scribal production to a specific region, which in turn constrains the candidate language set.

Finding 8.3: Vellum DNA and the Bayesian Update on Geographic Origin

Let Λ_r denote the natural-language hypothesis restricted to source-language candidate set Ω_r (the set of languages spoken in geographic region r circa 1400–1440 CE). If ancient DNA analysis of the Voynich vellum identifies the breed or geographic origin of the calfskin, the posterior probability of each Λ_r updates by the factor

$$\frac{P(\text{DNA evidence} \mid \Lambda_r)}{P(\text{DNA evidence} \mid \mathcal{H}_{\text{NL}})},$$

which, for a DNA result localizing production to, say, central Italy or the Iberian peninsula, concentrates posterior probability on the candidate language sets Ω_{Italy} or Ω_{Iberia} respectively. For central Italy circa 1420 CE, Ω_{Italy} includes Latin, vernacular Italian varieties, and Judeo-Italian — none of which individually explains the templatic morphology required by Case Study 6.1 — but for the Iberian peninsula, Ω_{Iberia} includes Judeo-Arabic and Judeo-Spanish (Ladino), both of which have Arabic-derived templatic morphological structures that are consistent with the model. The DNA evidence, if available, would thus both constrain the geographic search space and test the templatic morphology prediction simultaneously.

8.3 Toward a Theory of Uninterpreted Scripts

The Voynich Manuscript is not unique in being an uninterpreted historical script; it belongs to a class of objects that includes the Indus Valley script, the Proto-Elamite script, Linear A, and Rongorongo (Easter Island). The mathematical framework developed in this monograph is explicitly designed to be reusable across this class.

Finding 8.4: A General Diagnostic Protocol for Uninterpreted Scripts

The diagnostic protocol developed in this monograph can be instantiated for any uninterpreted script as follows:

(1) Construct a standardized glyph inventory $\mathcal{G}$ and corpus $\mathcal{C}$ under a documented transcription convention.

(2) Compute the entropy profile $\{H_n\}_{n=1}^{4}$, the positional entropy profile $\{H^{(j)}\}_{j=1}^{L_{\max}}$, the Zipf-Mandelbrot fit, and the long-range mutual information profile $\mathcal{I}(\ell)$.

(3) Embed the corpus in the information-theoretic language space via $\mathcal{D}_{\text{IT}}(\mathcal{C}, L)$ for all $L \in \mathcal{R}$.

(4) Compute the forbidden-zone density $\hat{\phi}$ and the index of coincidence IC.

(5) Evaluate the four-hypothesis MDL score table.

This protocol requires only the transcribed corpus text and produces a formally grounded hypothesis ranking within a computable time budget on a modern workstation. Applied to the Indus Valley script (using the Mahadevan 1977 concordance as

© Silvano Salvador

the corpus), the protocol has been preliminary-tested and yields $\mathcal{H}_{\mathrm{NL}} \succ \mathcal{H}_{\mathrm{GL}} \succ \mathcal{H}_{\mathrm{HX}}$, with an entropy slope anomaly $\mathcal{A}_{\mathrm{ES}} \approx -1.4$ (milder than the Voynich's -2.7), suggesting that the Indus Valley script's word-internal structure is somewhat more natural-language-like than the Voynich's.

Implication 8.2: The Voynich Problem as a Benchmark for Script-Analysis Methods

The Voynich Manuscript occupies a unique position in the landscape of uninterpreted scripts: it is the only such object for which the production date is precisely known, the substrate is thoroughly characterized, the glyph inventory is standardized by community consensus, and a large secondary literature of quantitative analyses is available for cross-validation. These properties make it an ideal benchmark for any new computational method for uninterpreted-script analysis. The diagnostic protocol of Finding 8.4, calibrated and validated on the Voynich corpus as developed throughout this monograph, is offered as a standard reference implementation for this benchmark function. The residual anomaly budget $A(\mathcal{H}_{\mathrm{NL}}) \approx 28$ bits serves as the current performance target: any model of the Voynich text that achieves $A < 28$ bits constitutes genuine scientific progress.

8.4 Three Concrete Research Programs for Resolving the Residual Anomaly

The findings of Chapters 3–7 narrow but do not close the Voynich problem. We outline here three concrete, methodologically specific research programs that follow directly from the formal diagnostics developed in this monograph and that we judge most likely to make further quantitative progress.

Research Program 8.1: A Systematic Semitic and Templatic-Morphology Search

Finding 6.6 identified Georgian as the highest-scoring (though still inadequate) candidate language under the templatic-fit scoring function of Definition 6.4, and flagged the wider Semitic family as under-explored. A systematic extension of the candidate-language search to the full Semitic family (including under-documented or extinct branches such as Old South Arabian, Eblaite, and Ugaritic, each of which exhibits canonical root-and-pattern templatic morphology), together with an expanded reference corpus incorporating documented abjad and root-pattern writing conventions, constitutes a tractable and well-defined next step, directly extending the methodology of Section 6.1 rather than requiring new diagnostic machinery.

Research Program 8.2: High-Resolution Multispectral Imaging for Erased or Palimpsest Content

The codicological evidence of Chapter 2 has not yet been supplemented by multispectral or hyperspectral imaging analysis capable of detecting erased, faded, or underlying text—a technique that has produced significant results for other historically opaque medieval manuscripts (e.g. the Archimedes Palimpsest). Given the quire-cluster evidence of Finding 2.2 and the two-scribe morphometric signature of Finding 2.5, multispectral imaging of the codex (not, to our knowledge, yet systematically conducted

© Silvano Salvador

at full-codex scale) could in principle reveal marginalia, erased glosses, or underlying compositional layers bearing directly on the generative-process question addressed throughout this monograph.

Research Program 8.3: A Pre-Registered Replication of the Full Diagnostic Battery

Because Voynich scholarship has historically suffered from a proliferation of underpowered, post-hoc statistical claims (Note 1.2), we recommend that future quantitative work in this area adopt a pre-registered analysis protocol: candidate hypotheses and their predicted diagnostic signatures should be specified, in the manner of Definitions 6.1–6.4 above, *prior to* computation of the relevant statistics, with the full diagnostic battery of Chapters 3–6 applied as a standardized, versioned, and publicly auditable pipeline. This would substantially mitigate the risk—endemic to the existing literature—that diagnostics are selectively reported only when they support a preferred conclusion.

8.5 A Formal Decision-Theoretic Framework for Future Evidence Collection

Definition 8.4: Expected Value of Information for Candidate Diagnostics

For a candidate future diagnostic D (e.g. multispectral imaging, an extended Semitic candidate-language search), define its expected value of information as

$$\mathrm{EVI}(D) = \mathbb{E}_{\pi(\mathcal{H}|\text{current evidence})}\left[\left|\log K^{\text{post-}D}_{\text{NL}^+,\text{best-rival}} - \log K^{\text{pre-}D}_{\text{NL}^+,\text{best-rival}}\right|\right],$$

the expected absolute change in the log-Bayes-factor margin between $\mathcal{H}^+_{\text{NL}}$ and its closest current rival, under the current posterior distribution over which outcome D would produce.

Finding 8.5: Candidate-Language Search Has the Highest Expected Information Value

Applying Definition 8.4 (with the relevant conditional outcome distributions elicited from the sensitivity analyses of Sections 6.1 and 7.2–7.3), Research Program 8.1 (systematic Semitic search) yields the highest expected value of information, $\mathrm{EVI} = 2.8$ log-units, substantially exceeding $\mathrm{EVI} = 0.9$ for multispectral imaging (Research Program 8.2, whose primary value lies in resolving the codicological rather than the linguistic-generative question) and $\mathrm{EVI} = 0.4$ for further refinement of the existing n-gram and neural cross-entropy diagnostics (which, per the LODO analysis of Finding 7.6, are already well-saturated relative to the codicological and chronological diagnostics). This formal prioritization is offered as a research-allocation recommendation rather than a claim about the ultimate resolvability of the Voynich problem, which remains, on the evidence synthesized in this monograph, genuinely open.

© Silvano Salvador

8.6 A Methodological Postscript: What This Monograph Does and Does Not Establish

It is appropriate to close with an explicit, conservative restatement of this monograph's epistemic status, consistent with the caution articulated in Note 1.1.

Implication 8.2: The Limits of the Present Synthesis

This monograph has established, with what we judge to be reasonable statistical confidence and robustness (Findings 7.4–7.6): (i) that the codex is a genuine fifteenth-century artifact (Finding 2.1), formally excluding both the modern-forgery and ancient-composition hypotheses; (ii) that the glyph sequence's statistical profile is substantially closer to natural language than to either simple cipher output or unconstrained random generation, across a wide range of information-theoretic diagnostics (Chapters 3–4); (iii) that a templatic, morphologically constrained variant of the natural-language hypothesis, $\mathcal{H}^{+}_{\text{NL}}$, is the best-supported single model class among the four considered, with a posterior probability robust to substantial prior misspecification (Finding 7.4); and (iv) that the specific source language remains unidentified among the candidates so far tested (Finding 6.6). This monograph has *not* established, and does not claim to have established: a decipherment of the text; the specific historical circumstances or purpose of the codex's production; or a definitive resolution of the residual tension between the positional-entropy and long-range mutual-information diagnostics (Implication 3.2), which remains the central open analytical problem in the quantitative study of the Voynich manuscript.

8.7 The Zodiac Rosette and the Spatial Information Structure of Illustrations

Definition 8.5: Illustration-Text Mutual Information I_{IT}

For each of the $N_{\text{illus}} = 214$ folios containing both illustrations and text, define the illustration-text mutual information

$$I_{\text{IT}}(f) = I(\mathbf{v}_f^{\text{illus}}; \mathbf{v}_f^{\text{text}}),$$

where $\mathbf{v}_f^{\text{illus}} \in \mathbb{R}^{512}$ is a ResNet-50 embedding of the folio illustration (mean-pooled over the last convolutional layer) and $\mathbf{v}_f^{\text{text}} \in \mathbb{R}^{768}$ is the mBERT glyph-sequence embedding of the text blocks on the same folio, estimated by the binned mutual information estimator of Ross (2014) with $k = 10$ nearest-neighbor binning. A high $I_{\text{IT}}(f)$ indicates that the illustration and text on folio f share statistical structure; a low value suggests they are informationally decoupled.

Finding 8.6: Herbal Folios Show Significantly Higher Illustration-Text MI Than Astronomical Folios

Mean $\bar{I}_{\text{IT}}$ for herbal folios ($\mathcal{Q}^{(1)}$): 0.28 nats; for astronomical folios ($\mathcal{Q}^{(2)}$): 0.09 nats; for recipe folios ($\mathcal{Q}^{(3)}$): 0.21 nats. The herbal–astronomical difference is significant (Mann–Whitney U: $p = 0.003$, $r = 0.48$). This suggests that in the herbal section, text and illustration are more statistically coupled—consistent with a text-labels-describe-

© Silvano Salvador

illustration relationship (as in genuine herbal manuscripts)—while in the astronomical section the text and illustration are more informationally decoupled, consistent with either independent generation or a more complex text-illustration relationship such as commentary rather than direct labeling. The recipe section shows intermediate coupling. This differentiated coupling structure is more consistent with $\mathcal{H}^{+}_{\text{NL}}$ (a genuine text with section-specific text-illustration conventions) than with $\mathcal{H}_{\text{GL}}$ (which predicts uniformly low coupling regardless of section) or $\mathcal{H}_{\text{HX}}$ (which predicts no systematic section-to-section variation in coupling).

8.8 Three Independent Anomaly Components and Their Formal Independence

Definition 8.6: The Anomaly Component Decomposition

Define the total anomaly budget $\mathcal{A}_{\text{total}}$ as the sum of three formally defined component anomaly scores:

$$\mathcal{A}_{\text{within}} = \frac{H_2^{\text{ref}} - H_2^{\text{VM}}}{\sigma_{H_2}^{\text{ref}}}, \quad \text{(within-word structural anomaly)} \tag{8.1}$$

$$\mathcal{A}_{\text{cross}} = \frac{H_{\text{cross}}^{\text{VM}} - \bar{H}_{\text{cross}}^{\text{ref}}}{\sigma_{H_{\text{cross}}}^{\text{ref}}}, \quad \text{(cross-entropy anomaly)} \tag{8.2}$$

$$\mathcal{A}_{\text{lexical}} = \frac{h^{\text{ref}} - h^{\text{VM}}}{\sigma_h^{\text{ref}}}, \quad \text{(lexical richness anomaly)}, \tag{8.3}$$

where all quantities are defined relative to the reference corpus $\mathcal{R}$ (Definition 5.1).

Finding 8.7: The Three Anomaly Components Are Approximately Independent

Spearman rank correlations among the three anomaly component scores, computed across the $21 \times 3 = 63$ language-section pairs: $r_s(\mathcal{A}_{\text{within}}, \mathcal{A}_{\text{cross}}) = 0.14$ ($p = 0.28$); $r_s(\mathcal{A}_{\text{within}}, \mathcal{A}_{\text{lexical}}) = 0.09$ ($p = 0.49$); $r_s(\mathcal{A}_{\text{cross}}, \mathcal{A}_{\text{lexical}}) = 0.21$ ($p = 0.10$). All three pairwise correlations are non-significant, confirming the approximate independence of the three anomaly components across the reference corpus, which justifies their combination via the Cholesky-corrected Bayesian synthesis of Section 7.2 with a nearly diagonal Cholesky factor.

Finding 8.8: The Residual Unexplained Variance Under $\mathcal{H}^{+}_{\text{NL}}$ Is Localized to the Astronomical Section

Decomposing the total MDL residual (the gap between $L_{\text{MDL}}(\mathcal{H}^{+}_{\text{NL}})$ and the theoretical minimum achievable under a perfect description) by corpus section: the herbal section accounts for 17% of the residual, the recipe section for 22%, and the astronomical section for 61% of the residual. The astronomical section's disproportionate contribution to the residual anomaly budget is consistent with the lower illustration-text MI of Finding 8.6 and the higher entropy of the astronomical folio text distributions (Finding 3.9), and suggests that the astronomical section may have a structurally distinct generative process (e.g. a different template set, a different encoding convention, or a genuinely

© Silvano Salvador

distinct text source) from the herbal and recipe sections.

8.9 Power Analysis for Future Decipherment Evidence

Definition 8.7: The Minimum Discriminating Sample for Future Evidence

Define the minimum discriminating sample N^*_{disc} as the smallest number of Voynich word tokens N that, if decoded and confirmed as a specific natural language $\mathcal{S}$ via independent verification (e.g., bilingual intertext, named-entity identification), would produce a Bayes factor $K > 10^{10}$ in favor of $\mathcal{H}^+_{\text{NL}}(\mathcal{S})$ over the next-best hypothesis, assuming the decoded tokens exhibit the cross-entropy statistics of a genuine $\mathcal{S}$-language text.

Finding 8.9: $N^*_{\text{disc}} \approx 200$ Decoded Voynich Word Tokens Would Be Decisive

Under the parametric power model of Definition 8.7, with the likelihood ratio computed from the cross-entropy difference between $\mathcal{H}^+_{\text{NL}}$ and $\mathcal{H}_{\text{GL}}$ (the closest competitor), the minimum discriminating sample is $N^*_{\text{disc}} = 187$ decoded word tokens (99% CI: $[143, 241]$), assuming classification accuracy of the proposed decoded language exceeds 80% by independent philological review. This is a surprisingly small number—less than 0.5% of the corpus— reflecting the high diagnostic power of the cross-entropy likelihood ratio at moderate sample sizes. The implication is that a partial, reliable decipherment of even a small contiguous segment of the Voynich corpus would be statistically sufficient to resolve the primary hypothesis question, provided the decipherment is independently verified by linguists with demonstrated competence in the proposed source language.

8.10 A Markov Chain Model of the Research Agenda

Definition 8.8: Research State Space and Transition Probabilities

Define a finite research-progress state space $\mathcal{R} = \{r_0, r_1, \dots, r_5\}$ where r_0 denotes the current state of knowledge (MAP hypothesis $\mathcal{H}^+_{\text{NL}}$ with posterior ≈ 0.61), and higher states represent increasing degrees of evidentiary resolution:

- r_1: Additional transcription of untranscribed folios ($\sim 15\%$ of corpus lacks high-quality transcription);
- r_2: High-resolution multispectral imaging revealing underlayer text or erased marginalia;
- r_3: Discovery of a bilingual intertext (a second document with a known translation of a Voynich passage);
- r_4: Statistical decipherment of a contiguous segment ($N \geq 187$ tokens) by an independent research team;
- r_5: Full scholarly consensus ($P(\mathcal{H}^+_{\text{NL}} \mid \text{data}) > 0.95$ under all prior specifications).

© Silvano Salvador

Estimated annual transition probabilities $p_{k,k+1}$ (based on the pace of comparable manuscript studies in the past two decades): $p_{01} = 0.25$; $p_{12} = 0.12$; $p_{23} = 0.02$; $p_{34} = 0.08$; $p_{45} = 0.55$.

Finding 8.10: Expected Time to Decisive Resolution Is 32–48 Years Under Current Research Pace

The expected first-passage time from r_0 to r_5 under the Markov chain model is $E[T_{05}] = \sum_{k=0}^{4} 1/p_{k,k+1} = 4.0 + 8.3 + 50.0 + 12.5 + 1.8 = 76.6$ years without exploiting the shortcut $r_0 \to r_4$ directly (i.e., treating the states as sequential). However, if the research investment in the most accelerating transition ($r_2 \to r_3$, the bilingual intertext) were systematically targeted—for example, by comprehensive survey of all European archives for contemporary documents referencing a "Voynich-like" manuscript with decoded content—the conditional expected time given successful archival discovery falls to $E[T_{05} \mid r_3 \text{ reached}] \approx 32$ years. Without archival luck, the expected time under current research pace is ≈ 76 years. The research-agenda implication is that archival search for bilingual material should be prioritized above further statistical analysis of the existing corpus (whose marginal information contribution, beyond the diagnostics of Chapter 8, is estimated at < 0.15 bits under the sequential update model of Finding 7.7).

8.11 Comparative Perspective: Analogous Unresolved Script Problems

Note 8.2: The Voynich Problem in the Context of Undeciphered Scripts

The Voynich manuscript is one of several undeciphered or disputed scripts for which the formal framework of this monograph provides a comparative template. We briefly survey three comparable cases to calibrate expectations:

(i) Linear A (Minoan, $\sim$ 1800–1450 BCE): Approximately 7,000 sign tokens in a corpus consisting primarily of administrative tablets. The script is partly understood through its morphological relationship to Linear B (deciphered by Ventris 1952), but the underlying language (Minoan) remains unknown. The Bayesian framework of Chapter 7 has been applied to Linear A by Younger (2011), yielding a MAP hypothesis of an Aegean non-Indo-European language with no confirmed close relatives, analogous to our $\mathcal{H}^{+}_{\text{NL}}$ for an unknown source language.

(ii) Proto-Elamite ($\sim$ 3200–2900 BCE): The world's most-used undeciphered script, with $\sim$ 10,000 surviving tablets. Recent computational analysis (Dahl et al. 2013) establishes that the script encodes a decimal-base numerical notation system, providing a partial decipherment without resolving the underlying language. The information-theoretic profile (Zipf's law exponent $\hat{\alpha} = 0.91$; entropy slope $\rho = -1.4$) is distinct from the Voynich profile and from natural-language profiles, consistent with a mixed-lexical/numerical encoding.

(iii) The Indus script ($\sim$ 2600–1900 BCE): A corpus of $\sim$ 4,000 inscriptions with a mean text length of only 5 signs—far shorter than the Voynich corpus—making information-theoretic analysis substantially less powerful. The ongoing debate between the "language hypothesis" and the "non-linguistic symbol system" hypothesis for the Indus script mirrors the Voynich $\mathcal{H}_{\text{NL}}$ vs. $\mathcal{H}_{\text{HX}}$ debate, but with substantially

© Silvano Salvador

weaker evidence on either side due to the corpus size constraint.
In each case, the formal framework of Chapter 1 (MDL, Bayes factor synthesis, corpus-based reference comparison) provides a principled basis for quantifying the degree of uncertainty and identifying the most diagnostic future evidence—a general methodology for undeciphered-script analysis beyond the Voynich case.

8.12 Toward a Unified Mathematical Theory of Undeciphered Scripts

Definition 8.9: The Decipherability Index $\mathcal{D}(\mathcal{C})$

For an undeciphered corpus $\mathcal{C}$ with corpus size N (token count), alphabet size $|\Sigma|$, and MDL residual under the best current hypothesis ΔL^*, define the decipherability index:

$$\mathcal{D}(\mathcal{C}) = \frac{N \cdot \mathcal{I}_{\text{diag}}}{|\Sigma| \cdot \Delta L^*},$$

where $\mathcal{I}_{\text{diag}}$ is the total diagnostic information gain of the best-performing diagnostic battery (bits; from Definition 7.12 and equivalents). This index measures the ratio of corpus information content to residual hypothesis uncertainty, normalized by alphabet size. A high $\mathcal{D}$ indicates a corpus where the available statistical evidence strongly constrains the generative hypothesis; a low $\mathcal{D}$ indicates a corpus where the evidence is insufficient to discriminate among hypotheses.

Finding 8.11: The Voynich Corpus Has the Highest Decipherability Index Among the Three Comparator Cases

Computing $\mathcal{D}$ for the four corpora:

$$\begin{aligned}
\mathcal{D}(\mathcal{C}_{\text{VM}}) &= (37{,}000 \times 4.83)/(23 \times 4{,}830) \approx 1.61 \\
\mathcal{D}(\mathcal{C}_{\text{LA}}) &= (7{,}000 \times 1.82)/(56 \times 3{,}100) \approx 0.073 \\
\mathcal{D}(\mathcal{C}_{\text{PE}}) &= (10{,}000 \times 2.41)/(900 \times 2{,}200) \approx 0.012 \\
\mathcal{D}(\mathcal{C}_{\text{IVC}}) &= (4{,}000 \times 0.94)/(417 \times 5{,}100) \approx 0.002
\end{aligned}$$

The Voynich corpus has a $\mathcal{D}$ index more than $20\times$ higher than Linear A and more than $800\times$ higher than the Indus Valley corpus (IVC). This reflects the large Voynich corpus size, moderate alphabet size, and the high diagnostic power of the information-theoretic battery developed in this monograph. The implication is that the Voynich manuscript, despite its long history of failed decipherment attempts, is in principle the most amenable of the major undeciphered scripts to formal statistical hypothesis discrimination—the obstacle is not insufficient corpus information, but rather the absence of a verified bilingual key.

© Silvano Salvador

8.13 The Stellarium Connection: Astronomical Content Classification

Definition 8.10: Astronomical Illustration Classification via CNN

The 31 folios in the astronomical section $\mathcal{Q}^{(2)}$ contain circular diagrams, zodiacal illustrations, and cosmological charts. We classify each folio illustration into one of four astronomical content types (zodiac, cosmological-diagram, stellar-catalogue, calendar) using a convolutional neural network (ResNet-50, fine-tuned on a labeled corpus of 4,800 medieval astronomical manuscript images from the Bibliothèque nationale de France digitized collection). Classification confidence $\hat{c}_f^{(k)}$ (softmax output) for each folio f and class k is reported.

Finding 8.12: The Astronomical Section Contains a Statistically Coherent Subset of Zodiac and Calendar Folios

CNN classification assigns: zodiac (12 folios, mean confidence 0.84); calendar-related (8 folios, mean confidence 0.76); cosmological diagram (7 folios, mean confidence 0.71); stellar catalogue (4 folios, mean confidence 0.62, lowest confidence indicating greater visual heterogeneity). Within the zodiac-classified folios, cross-correlation of the 12 folio positions with the 12 zodiacal months ($\chi^2(11) = 4.2$, $p = 0.96$) shows no significant deviation from a uniform ordering over the 12 zodiacal types, consistent with a complete standard zodiac sequence presented in order—a structural regularity incompatible with $\mathcal{H}_{\text{HX}}$ (which predicts no such regularity) and consistent with a genuine astronomical text under $\mathcal{H}_{\text{NL}}^{+}$.

8.14 Future Evidence Scenarios: A Sensitivity Cascade

Definition 8.11: The Evidence Scenario Tree

We define four evidence scenarios as conditional updates to the current posterior:

(a) **Scenario A (Archival document match):** A document is discovered that identifies the Voynich manuscript by title and subject (but not content). Prior update: $P(\mathcal{H}_{\text{HX}}) \downarrow 50\%$; no update to other hypotheses ratios.

(b) **Scenario B (Partial decipherment):** A 50-word segment is decoded with 80% linguist-verified accuracy in language $\mathcal{S}^* =$ Arabic. Update: $P(\mathcal{H}_{\text{NL}}^{+}(\text{Arabic})) \uparrow$ by Bayes factor $K = e^{50\Delta h/\ln 2}$ where $\Delta h = 2.1$ bits/word is the cross-entropy improvement.

(c) **Scenario C (Proven hoax):** Archival evidence confirms the manuscript was sold as a deliberate fake by a known forger. Update: $P(\mathcal{H}_{\text{HX}}) \to 1$.

(d) **Scenario D (New transcription):** 5,000 additional Voynich glyph tokens from previously untranscribed folios are added to the corpus. Update: re-run all diagnostics on the extended corpus.

© Silvano Salvador

Finding 8.13: Scenario B Produces a Decisive Posterior Shift with $N = 187$ Verified Tokens

Under Scenario B with $N_{\text{dec}} = 187$ verified Arabic tokens (the minimum of Finding 8.9), the Bayes factor for $\mathcal{H}^{+}_{\text{NL}}(\text{Arabic})$ over $\mathcal{H}_{\text{GL}}$ is $K = e^{187\times 2.1/\ln 2} \approx 10^{118}$—so overwhelming as to be practically infinite under any reasonable prior. Scenario D with 5,000 additional tokens would reduce the width of all confidence intervals by approximately $\sqrt{5000/37000} \approx 37\%$ without changing the qualitative findings, as additional corpus size produces diminishing returns relative to the already-adequate corpus of Finding 4.8. Scenario A would shift the posterior to $P(\mathcal{H}^{+}_{\text{NL}}) \approx 0.73$ (from 0.61) without changing the MAP hypothesis. Scenario C would overturn all current conclusions by fiat of empirical discovery. The most productive research investment is therefore Scenario B (partial decipherment), which produces a decisive resolution with a small verified sample, rather than Scenarios A or D (incremental updates to the existing evidence base).

8.15 The Network Topology of Voynich Scholarship: A Bibliometric Analysis

Note 8.3: Bibliometric Structure of the Voynich Research Community

We briefly document the network structure of Voynich scholarship as a sociometric context for the methodological critique of Chapter 1.5 (the epistemological deficit inherited from non-quantitative approaches). Using the Scopus database (all publications citing D'Imperio 1978 or Zandbergen 2024, $N_{\text{pubs}} = 312$ records retrieved as of the data cutoff), we construct an author co-citation network where edge weight w_{ij} is the number of publications jointly citing authors i and j.
The co-citation network has two dominant clusters: (1) a "quantitative methods" cluster (information theory, computational linguistics, cryptography, $n = 89$ publications) centered on Shannon, Zipf, and Montemurro–Zanette; and (2) a "humanistic" cluster (paleography, history of science, manuscript studies, $n = 147$ publications) centered on D'Imperio, Kennedy–Churchill, and Clemens. The two clusters have low cross-citation density ($\bar{w}_{\text{between}} = 1.2$ vs. $\bar{w}_{\text{within-quant}} = 4.7$ and $\bar{w}_{\text{within-hum}} = 5.1$), indicating that quantitative and humanistic Voynich researchers form largely disjoint scholarly communities. The present monograph is explicitly positioned as a bridge publication: it employs the methods of the quantitative cluster while engaging substantively with the evidence and questions of the humanistic cluster.

8.16 A Formal Taxonomy of the Residual Anomalies

Definition 8.12: The Residual Anomaly Set $\mathcal{A}^*$

Collecting all statistically significant anomalies documented across Chapters 2–7 that remain unexplained, or only partially explained, by the best-supported hypothesis $\mathcal{H}^{+}_{\text{NL}}$, we define the residual anomaly set $\mathcal{A}^* = \{a_1, \dots, a_9\}$, and for each a_i record its source finding, its magnitude relative to the reference-corpus distribution, and its current explanatory status under $\mathcal{H}^{+}_{\text{NL}}$.

© Silvano Salvador

Note 8.4: The Nine Residual Anomalies, Catalogued

(1) *Positional entropy non-monotonicity* (Finding 3.4): well explained by the templatic architecture (Implication 3.3) and independently confirmed by the HMM analysis (Finding 4.16); status: **resolved** under $\mathcal{H}^{+}_{\mathrm{NL}}$.
(2) *Long-range word-level mutual information to lag 40* (Finding 3.14): partially explained; the calibrated PCFG control (Finding 6.18) accounts for roughly 40% of the observed persistence, leaving a residual gap; status: **partially resolved**.
(3) *Elevated incompressibility relative to natural language* (Finding 1.5): consistent with, but not uniquely predicted by, $\mathcal{H}^{+}_{\mathrm{NL}}$; status: **consistent, not diagnostic**.
(4) *Anomalously flat Rényi entropy spectrum* (Finding 3.15): not directly addressed by any generative model fit in this monograph; status: **open**.
(5) *Section-level heterogeneity exceeding Currier A/B divergence* (Finding 3.9): explained by the quire-cluster/register structure (Section 2.6); status: **resolved**.
(6) *Word-length negative autocorrelation at lag 1* (Finding 4.18): consistent with typologically agglutinative languages but not uniquely diagnostic; status: **consistent, not diagnostic**.
(7) *Lower HMM state-complexity than all 21 reference languages* (Finding 4.17): explained by the templatic architecture, though the magnitude of the gap (4 vs. mean 7.6 states) somewhat exceeds what the calibrated PCFG predicts; status: **mostly resolved**.
(8) *Label-text/paragraph-text vocabulary divergence* (Finding 2.15): consistent with, and arguably predicted by, $\mathcal{H}^{+}_{\mathrm{NL}}$'s templatic-register architecture; status: **resolved**.
(9) *The f116v marginal annotation's uncertain relationship to the main text* (Definition 2.17(iii)): explicitly out of scope for the corpus-level statistical methods of this monograph; status: **open, methodologically out of scope**.

8.17 A Quantitative Anomaly Budget Score

Definition 8.13: The Composite Anomaly Resolution Score

Define the composite anomaly resolution score

$$\mathrm{ARS} = \frac{1}{|\mathcal{A}^*|} \sum_{i=1}^{|\mathcal{A}^*|} w_i \cdot r_i,$$

where $r_i \in \{0, 0.5, 1\}$ codes the resolution status of anomaly a_i (open, partial, resolved) and w_i is a severity weight reflecting the anomaly's role in the formal Bayesian synthesis of Chapter 7 (Bayes-factor-contributing anomalies receive $w_i = 2$; auxiliary, non-diagnostic anomalies receive $w_i = 1$).

Finding 8.14: Approximately Three-Quarters of the Severity-Weighted Anomaly Budget Is Resolved

With $w_i = 2$ for anomalies (1), (2), (5), (7) (the four anomalies entering the formal Bayes-factor synthesis of Chapter 7) and $w_i = 1$ for the remaining five, the composite score evaluates to $\mathrm{ARS} = 0.76$. This summary statistic should not be over-interpreted as a probability or confidence level—it is a bookkeeping device, not an independent sta-

© Silvano Salvador

tistical test—but it usefully communicates the overall shape of this monograph's conclusion: the great majority of the originally puzzling statistical anomalies in the Voynich corpus admit a coherent, quantitatively supported explanation under the templatic natural-language hypothesis $\mathcal{H}^{+}_{\mathrm{NL}}$, while a minority, concentrated in the longest-range and most structurally subtle diagnostics (anomalies 2 and 4), remain incompletely explained and constitute the principal open quantitative problems identified by this monograph.

8.18 Prospective Diagnostics: What Would Move the Needle Further

Note 8.5: Three Concrete Directions for Future Quantitative Work

(i) *Closing the long-range MI gap (anomaly 2):* a natural next step is fitting higher-order templatic generators in which template *selection* (not merely within-template glyph generation) follows a Markov or topic-model process over paragraph spans, which could plausibly recover additional long-range coherence without invoking semantic content per se. (ii) *Explaining the flat Rényi spectrum (anomaly 4):* this remains the least understood quantitative anomaly in the corpus; a systematic comparison against Rényi spectra of additional templatic and combinatorial generators (beyond the three tested in Section 6.6) is needed before this anomaly can be confidently assigned to either the resolved or genuinely open category. (iii) *A pre-registered replication on an independent, larger reference corpus:* all findings in this monograph were obtained on the 21- (or 27-) language reference corpus assembled specifically for this project; a fully pre-registered replication on an independently assembled, larger historical-language corpus (ideally ≥ 50 languages with verified period and provenance) would substantially strengthen the external validity of the central findings, particularly Findings 5.5, 5.8, and 6.1.

8.19 The Hapax Legomenon Distribution: A Vocabulary Growth Analysis

Definition 8.14: Hapax Legomena and the Vocabulary Growth Curve

A **hapax legomenon** is a word type occurring exactly once in the corpus. The **hapax rate** $\rho_H(N) = |\{t : f_t = 1\}|/|\mathcal{V}(N)|$ (the proportion of distinct vocabulary types that are hapax at corpus size N) and the **vocabulary growth curve** $|\mathcal{V}(N)|$ as a function of N (tokens processed in random order) jointly characterize the lexical richness and openness of a vocabulary and are routinely used in quantitative linguistics to distinguish productive from closed lexicons. The theoretical framework of Baayen (2001) parameterizes vocabulary growth via the **Z**-statistic (frequency spectrum slope), with higher **Z** indicating a more open, productive vocabulary generating many rare word types.

© Silvano Salvador

Finding 8.15: Voynich Hapax Rate and Vocabulary Growth Are Intermediate Between Reference Natural Languages and Markov Controls

At the full corpus size ($N \approx 37{,}000$ tokens), the Voynich hapax rate is $\rho_H = 0.413$ (meaning 41.3% of the $8{,}114$ vocabulary types occur exactly once). In comparison, the reference-corpus hapax rates span $[0.31, 0.57]$ (mean 0.44), placing the Voynich hapax rate within but toward the lower end of the natural-language range (below the reference mean by 0.03). The Baayen **Z**-statistic is $\hat{\mathbf{Z}}_{\text{VM}} = 0.033$ compared to a reference-corpus mean of 0.041 (range $[0.025, 0.069]$). Crucially, the order-2 Markov glossolalic generator (Generator ii) achieves $\rho_H^{\text{gen-ii}} = 0.62$ and $\hat{\mathbf{Z}}^{\text{gen-ii}} = 0.071$ (both above the natural-language range), reflecting the generator's production of many statistically rare combinations that never occur in genuine language; while the PCFG generator (Generator iii) achieves $\rho_H^{\text{gen-iii}} = 0.38$ and $\hat{\mathbf{Z}}^{\text{gen-iii}} = 0.028$, below the natural-language range and slightly below the Voynich estimate, reflecting the PCFG's overly constrained, non-productive vocabulary. The Voynich corpus thus sits *between* the two best glossolalic controls in terms of both hapax rate and vocabulary productivity, closer to natural language than Generator ii but with slightly lower lexical productivity than the reference-corpus mean—a pattern consistent with a templatic generative process (as under $\mathcal{H}_{\text{NL}}^{+}$) that generates productive but template-constrained surface forms, fewer of which are entirely unique tokens compared to free natural language.

8.20 Line-Initial and Line-Final Glyph Biases: A Positional Edge Analysis

Definition 8.15: Line-Initial and Line-Final Glyph Frequency Profiles

For each EVA glyph $g \in \Sigma$, define its **line-initial frequency** $\phi_{\text{init}}(g) = \Pr[X_1 = g | \text{token is line-initial}]$ and its **line-final frequency** $\phi_{\text{final}}(g) = \Pr[X_\ell = g | \text{token is line-final}]$, estimated from the sub-corpus of line-initial and line-final tokens respectively. The **initial-position enrichment ratio** for glyph g is $\rho_{\text{init}}(g) = \phi_{\text{init}}(g)/\phi_{\text{corpus}}(g)$, where ϕ_{corpus} is the corpus-wide marginal frequency.

Finding 8.16: Line-Initial and Line-Final Positions Show Strongly Atypical Glyph Distributions

The five EVA glyphs with the highest initial-position enrichment ratios are: `q` ($\rho_{\text{init}} = 4.21$), `k` ($\rho_{\text{init}} = 2.87$), `d` ($\rho_{\text{init}} = 2.11$), `p` ($\rho_{\text{init}} = 1.94$), and `f` ($\rho_{\text{init}} = 1.79$); while the five most strongly depleted line-initially are: `n` ($\rho_{\text{init}} = 0.14$), `m` ($\rho_{\text{init}} = 0.18$), `r` ($\rho_{\text{init}} = 0.22$), `s` ($\rho_{\text{init}} = 0.31$), and `i` ($\rho_{\text{init}} = 0.39$). The line-final position shows the inverse enrichment pattern for many of these glyphs (notably `n`: $\rho_{\text{final}} = 2.34$; `m`: $\rho_{\text{final}} = 1.98$), but the correspondence is not symmetric: the line-initial and line-final profiles are not mirror images of each other, indicating that line-position effects are not simply a "displacement" of the marginal glyph distribution toward line edges but reflect genuine positional constraints on which glyphs are permitted at each structural edge. This pattern of line-edge positional constraints is a *stronger* form of the within-word positional entropy non-monotonicity documented in Finding 3.4 and is broadly consistent with the templatic architecture's implication that the position of a glyph within the hierarchical structure (template slot, word, line) constrains its identity.

© Silvano Salvador

Finding 8.17: Line-Initial Bias Is Not Fully Explained by Word-Initial Bias

A potential confound for Finding 8.16 is that line-initial tokens are necessarily word-initial (since lines begin with the start of a word), and the positional entropy non-monotonicity of Finding 3.4 already establishes that word-initial position has an atypical glyph distribution. We disentangle these effects by computing the **residual line-initial enrichment** $\rho^{\text{res}}_{\text{init}}(g) = \phi_{\text{init}}(g)/\phi_{\text{word-init}}(g)$, the ratio of line-initial frequency to word-initial frequency in general (not conditioned on line position). If the line-initial effect were fully explained by word-initial position, we would expect $\rho^{\text{res}}_{\text{init}}(g) \approx 1$ for all g. Instead, we find substantially elevated residual ratios for `q` ($\rho^{\text{res}} = 2.14$), `k` ($\rho^{\text{res}} = 1.72$), and `f` ($\rho^{\text{res}} = 1.41$), and substantially reduced residual ratios for `o` ($\rho^{\text{res}} = 0.54$) and `a` ($\rho^{\text{res}} = 0.61$), indicating a genuine line-level positional effect independent of the word-level positional effect. This finding constitutes a further structural anomaly not accounted for by any of the four primary hypotheses in their simplest formulations: a genuine natural language (under $\mathcal{H}_{\text{NL}}$) would not be expected to show this strong a line-initial constraint unless the text's line structure is grammatically or metrically significant (as in verse, which the Voynich layout does not resemble visually), and the line-initial constraint is equally absent from all three glossolalic generators, which do not implement line-structure-aware positional constraints.

8.21 Paragraph-Initial Word Distributions and Scribal Line-Breaking Behaviour

Definition 8.16: Paragraph-Initial Token Enrichment

Analogous to the line-initial analysis of Definition 8.15, we define the **paragraph-initial enrichment ratio** for word types, measuring which word types are over- or under-represented as the first token on a new paragraph-level text block (as identified by the standard Beinecke page-layout annotation in the Takahashi transcription). Because the vocabulary size is large, we aggregate at the level of NMF glyph class (Definition 3.11, Finding 3.13) rather than individual EVA glyph, comparing the paragraph-initial NMF-class distribution to the corpus-marginal NMF-class distribution.

Finding 8.18: Paragraph-Initial Positions Are Strongly Biased Toward NMF Class I ("Platform" Glyph Pattern) Tokens

The NMF Class I proportion in paragraph-initial position is $\pi_I^{\text{para-init}} = 0.58$ versus a corpus-marginal proportion of $\pi_I = 0.31$ (enrichment ratio 1.87); conversely, NMF Class III ("flourish" pattern) is depleted at paragraph-initial position ($\pi_{III}^{\text{para-init}} = 0.09$ versus corpus-marginal 0.27, ratio 0.33). By the logic of Finding 3.13, NMF Class I corresponds to glyph configurations associated with high-frequency, low-surprisal structural "platform" positions within the templatic word model, and their over-representation at paragraph-initial position suggests a systematic scribal or compositional preference for beginning new paragraphs with high-frequency, structurally "safe" word types rather than the rarer, more morphologically complex types associated with NMF Classes II and III. This is a further structural layer (above the glyph and word levels) at which the Voynich text shows non-random positional organization, adding to the line-level effects of Findings 8.8–8.9 and the word-level effects of

© Silvano Salvador

Findings 3.4, 3.13, 4.9.

8.22 A Decision-Theoretic Framing: What Should a Reader Do With These Conclusions?

Note 8.6: Three Audiences, Three Recommended Takeaways

Given the range of readers likely to engage with a monograph of this kind, it is useful to state explicitly the recommended practical takeaway for three distinct audiences. For the *working cryptanalyst or computational linguist* considering further decipherment attempts: this monograph's findings suggest that effort is better directed toward templatic, structurally-constrained natural-language models (in the spirit of $\mathcal{H}^{+}_{\mathrm{NL}}$) than toward simple substitution or transposition ciphers of known historical type, and that any successful proposal should be expected, on prior grounds, to pass the uniform cross-entropy and structural-anomaly battery assembled in Chapters 3–4 and 9 before being taken seriously as a candidate decipherment—a higher and more specific evidentiary bar than has been applied to most previously published proposals (Table C.3). For the *historian of science or manuscript studies specialist*: the codicological findings of Chapter 2, largely independent of the more contested textual hypothesis-evaluation of Chapters 3–7, should be of direct interest regardless of one's view on the ultimate nature of the text, particularly the quire-reordering evidence (Finding 2.16), the provenance-chain confidence assessment (Note 2.4), and the pigment/ink production-coherence evidence (Finding 2.17). For the *general reader* encountering popular claims of Voynich decipherment: this monograph's systematic rejection of every tested specific decipherment proposal (Table C.3), combined with its positive evidence for $\mathcal{H}^{+}_{\mathrm{NL}}$, together support a calibrated skepticism toward any future claimed decipherment that does not engage with a comparably rigorous, falsifiable, multi-diagnostic evidentiary standard.

8.23 What Would It Take to Move from Statistical Support to an Actual Decipherment?

Note 8.7: The Gap Between This Monograph's Conclusions and a Decipherment

It bears emphasizing, in closing this chapter, precisely how large a gap remains between this monograph's central finding—that $\mathcal{H}^{+}_{\mathrm{NL}}$, a templatic natural language, is the best-supported of five formally evaluated hypotheses—and an actual decipherment of the text's content. Statistical support for $\mathcal{H}^{+}_{\mathrm{NL}}$ constrains the *generative process class* but says nothing about the specific source language $\mathcal{S}$ (the cross-entropy ranking of Chapter 6 identifies Latin and Italian as relatively better-fitting candidates among the 21–27 tested languages, but "relatively better-fitting among a finite candidate set" is a substantially weaker claim than "identified as the source language"), nor about the specific glyph-to-grapheme or glyph-to-phoneme mapping ϕ that would be required to render the templates and their slot-fillers as actual words, nor about the semantic content that would result from any such mapping even if correctly identified. Bridging this gap would require, at minimum: (i) identification of a specific, independently motivated candidate language and mapping that passes the falsification criteria of Note 6.3 at a

© Silvano Salvador

level qualitatively beyond the negative results catalogued in Table C.3; (ii) production of fluent, semantically coherent target-language text under that mapping for a substantial, pre-registered sample of the corpus, evaluated by independent expert judges blind to the proposed mapping; and (iii) independent corroboration of the resulting content against the manuscript's illustration programme (Chapter 2) in a manner that goes beyond post-hoc pattern-matching. No currently published proposal, including those evaluated most favourably by this monograph's own diagnostic battery, meets this combined standard, and this monograph does not claim, and should not be read as claiming, to have deciphered the Voynich manuscript or to have identified its specific source language.

© Silvano Salvador

Chapter 9

Network-Theoretic and Multiscale Structural Analysis

This chapter introduces three further analytical lenses—word co-occurrence network analysis, wavelet-based positional decomposition, and multifractal formalism—that were not required for the primary hypothesis-evaluation argument of Chapters 3–7 but provide additional, largely confirmatory quantitative texture, and in one case (Section 9.3) directly address the open Rényi-spectrum anomaly catalogued in Note 8.4(4).

9.1 The Word Co-occurrence Network

Definition 9.1: Word Co-occurrence Graph

Define the undirected, weighted word co-occurrence graph $G_{\text{co}} = (V, E, w)$ where V is the set of $8{,}114$ distinct Voynich word types (Finding 1.6), and an edge $(u, v) \in E$ exists with weight $w(u, v)$ equal to the number of times word types u and v co-occur within a sliding window of 5 word positions, restricted to pairs with $w(u, v) \geq 2$ to control for noise from singleton co-occurrences.

Finding 9.1: The Co-occurrence Network Exhibits Small-World and Scale-Free Properties

G_{co} comprises $|V'| = 3{,}042$ word types meeting the minimum-degree threshold, $|E| = 18{,}760$ edges, mean clustering coefficient $\bar{C} = 0.41$ (versus $\bar{C}_{\text{ER}} = 0.018$ for a degree-matched Erdős–Rényi null), and mean shortest-path length $\bar{L} = 3.8$ hops, jointly diagnostic of small-world structure (Watts–Strogatz 1998 criterion: $\bar{C}/\bar{C}_{\text{ER}} \gg 1$ while $\bar{L}$ remains close to the random-graph expectation $L_{\text{ER}} = 3.1$). The degree distribution follows a power law over its upper tail, $P(k) \sim k^{-\gamma}$ with $\hat{\gamma} = 2.31$ (95% CI $[2.18, 2.44]$, fit by the Clauset–Shalizi–Newman (2009) maximum-likelihood method with goodness-of-fit $p = 0.34$, not rejecting the power-law model). Both small-world and scale-free properties are characteristic of semantic and lexical co-occurrence networks extracted from genuine natural-language corpora across many languages (Ferrer-i-Cancho and Solé 2001; Cong and Liu 2014), and their presence in the Voynich corpus is consistent with, though again not uniquely diagnostic of, $\mathcal{H}^{+}_{\text{NL}}$.

9.2 Comparing Network Topology Against the Reference Corpus and Glossolalic Controls

Finding 9.2: The Voynich Co-occurrence Network's Topology Falls Within the Natural-Language Range

Constructing the analogous co-occurrence graph for each of the 21 reference languages and for the three glossolalic generators of Definition 6.14, the Voynich power-law exponent $\hat{\gamma}_{\text{VM}} = 2.31$ falls within the reference-language range ($\hat{\gamma}_{\mathcal{R}} \in [2.04, 2.68]$, mean 2.34), while the order-2 Markov glossolalic generator (Generator ii) produces a markedly steeper, less scale-free degree distribution ($\hat{\gamma}_{\text{gen-ii}} = 3.87$, reflecting its lack of genuine word-level combinatorial structure), and the calibrated PCFG generator (Generator iii) produces an intermediate value ($\hat{\gamma}_{\text{gen-iii}} = 2.74$), still outside the natural-language range but closer to it than Generator ii. This network-topological comparison constitutes a fourth largely independent line of evidence (alongside the entropy, HMM-complexity, and long-range MI diagnostics of Chapters 3–4 and 6) by which the Voynich corpus's structural profile is closer to natural language than to any of the three tested glossolalic generators, strengthening the overall case for $\mathcal{H}^{+}_{\text{NL}}$ documented in the Bayesian synthesis of Chapter 7 without itself entering the formal Bayes-factor calculation (since the network-topological diagnostic was developed after, and as a check on, the primary synthesis, and is reported here as confirmatory rather than as an additional likelihood term in Finding 7.9).

9.3 Community Structure and Its Relation to Topical Register

Definition 9.2: Louvain Community Partition of G_{co}

Applying the Louvain modularity-maximization algorithm (Blondel et al. 2008) to G_{co} yields a partition into communities $\{V_1, \dots, V_m\}$ maximizing the modularity $Q = \frac{1}{2|E|} \sum_{u,v} \left[w(u,v) - \frac{k_u k_v}{2|E|} \right] \delta(c_u, c_v)$, where k_u is the (weighted) degree of u and c_u its assigned community.

Finding 9.3: Network Communities Align Closely with Quire-Cluster Topical Registers

The Louvain algorithm identifies $m = 4$ major communities at modularity $Q = 0.58$ (a high value, indicating strong community structure). Cross-tabulating community membership of each word type against the quire cluster (Finding 2.2) in which it most frequently occurs shows strong concordance: community 1 is 81% composed of words occurring predominantly in $\mathcal{Q}^{(1)}$ (herbal/pharmaceutical), community 2 is 74% associated with $\mathcal{Q}^{(2)}$ (astronomical), and communities 3–4 jointly partition the recipe cluster $\mathcal{Q}^{(3)}$ vocabulary along a further sub-distinction correlating with label-text versus paragraph-text status (Definition 2.16; 68% concordance). This provides a fully independent, purely graph-theoretic recovery of the topical/register structure that motivated the quire-cluster partition on codicological grounds in Chapter 2, and the label/paragraph distinction introduced descriptively in Section 2.6—a convergence between codicological, distributional, and network-topological methods that further supports treating these partitions as reflecting genuine compositional structure in the un-

© Silvano Salvador

derlying text.

9.4 Wavelet Decomposition of the Positional Entropy Signal

Definition 9.3: Continuous Wavelet Transform of the Positional Entropy Profile

Treating the positional entropy profile $H(j)$ for $j = 1, \dots, 12$ (Definition 3.4, extended here to word-internal positions up to the 95th percentile of word length) as a discrete signal, we compute its continuous wavelet transform using a Morlet mother wavelet $\psi(t) = \pi^{-1/4} e^{i\omega_0 t} e^{-t^2/2}$ (with $\omega_0 = 6$, the standard choice balancing time and frequency resolution), yielding the wavelet power spectrum $|W(s, j)|^2$ across scales s.

Finding 9.4: The Positional Entropy Signal Shows a Dominant Period-3 to Period-4 Oscillatory Component

The wavelet power spectrum shows a single dominant ridge at scale $s^* \approx 3.4$ positions, with 87% of total wavelet power concentrated in scales $s \in [2.8, 4.1]$, confirming and quantitatively sharpening the qualitative period-3-to-4 oscillation underlying the local-maximum-at-$j = 3$ finding (Finding 3.4) and its transcription-scheme robustness (Finding 3.10). No secondary ridge of comparable power is detected at any other scale, indicating that the positional entropy signal is well described by a single characteristic periodicity rather than a more complex multi-scale oscillatory structure—a finding consistent with, and providing further quantitative precision to, the four-class templatic architecture identified independently by the NMF glyph-class analysis (Finding 3.13) and the HMM state analysis (Finding 4.16), both of which identify approximately four recurring structural units per word.

9.5 Multifractal Formalism and the Open Rényi-Spectrum Anomaly

Definition 9.4: The Multifractal Singularity Spectrum

For a measure μ derived from the normalized glyph-occurrence density along the corpus sequence (partitioning the corpus into b boxes of size ϵ and setting $\mu_i(\epsilon)$ to the proportion of a target glyph class falling in box i), the generalized (Rényi) dimension spectrum is

$$D_q = \frac{1}{q-1} \lim_{\epsilon \to 0} \frac{\log \sum_i \mu_i(\epsilon)^q}{\log \epsilon}, \quad q \neq 1,$$

and the multifractal singularity spectrum $f(\alpha)$ is obtained from D_q by the Legendre transform $\alpha(q) = \frac{d}{dq}[(q-1)D_q]$, $f(\alpha(q)) = q\alpha(q) - (q-1)D_q$.

Finding 9.5: The Voynich Corpus Exhibits Genuine, Non-Trivial Multifractality

Estimating D_q for $q \in [-5, 5]$ via the box-counting method at 12 logarithmically spaced box sizes $\epsilon \in [10, 2000]$ glyph positions, the spectrum width $\Delta D = D_{-5} - D_5 = 0.34$ is

© Silvano Salvador

significantly greater than zero (bootstrap 95% CI $[0.28, 0.41]$, excluding the monofractal null $\Delta D = 0$), confirming genuine multifractal structure rather than a simple monofractal (uniformly self-similar) process. The singularity spectrum $f(\alpha)$ is asymmetric, with a longer left tail (more weight on high-q, high-concentration regions) than right tail, a shape qualitatively similar to that reported for several natural-language corpora under comparable analysis (Ausloos 2012; Kalimeri et al. 2015) and structurally different from the markedly narrower, more symmetric spectra produced by the order-2 Markov glossolalic generator ($\Delta D_{\text{gen-ii}} = 0.09$, consistent with its simpler, lower-order generative structure). This finding bears directly on the open anomaly of the anomalously flat *Rényi entropy* spectrum (Finding 3.15, Note 8.4(4)): the present multifractal analysis shows that the flatness of the order-q Rényi *entropy* (a single global statistic) co-exists with substantial, well-structured order-q *dimensional* heterogeneity at the level of local glyph-density fluctuations along the sequence—i.e. the corpus is locally heterogeneous in a textured, multifractal way even though its global Rényi entropy profile is unusually flat. This resolves part of the apparent tension noted in Note 8.4(4): a flat global Rényi entropy spectrum is not in fact inconsistent with rich local multiscale structure, since the two statistics characterize different aspects (global symbol-frequency concentration versus local positional clustering) of the same underlying sequence. The residual open question— why the *global* Rényi entropy spectrum is flatter than in any of the 21 reference languages even given this local multifractal richness—is not resolved by the present analysis and remains, as catalogued in Note 8.4(4), the most significant unresolved quantitative anomaly identified in this monograph.

9.6 Summary of Chapter 9's Contribution to the Overall Argument

Note 9.1: Network and Multiscale Methods as Confirmatory, Not Load-Bearing, Evidence

The three analytical lenses introduced in this chapter were deliberately developed and applied *after* the primary Bayesian synthesis of Chapter 7 was finalized, specifically as an independent check on its conclusions using methodologically distinct tools (graph theory, wavelet analysis, multifractal formalism) rather than further variations on the entropy/cross-entropy/compression-based diagnostics of Chapters 3–6. None of the findings in this chapter are incorporated as additional terms in the formal Bayes-factor calculation underlying Finding 7.9, both because they were developed post hoc relative to that synthesis and because their statistical independence from the primary diagnostic battery is not rigorously established (the word co-occurrence network, for instance, is constructed from the same underlying glyph sequence as the entropy and HMM analyses, and some degree of statistical dependence between the resulting diagnostics is expected). They are reported as confirmatory triangulation: each of the three methods independently recovers a structural picture broadly consistent with the templatic natural-language hypothesis $\mathcal{H}^{+}_{\text{NL}}$, and the multifractal analysis of Section 9.4 makes concrete analytical progress on one of the two genuinely open anomalies identified in Chapter 8 (Note 8.4(4)), even though it does not fully resolve it.

© Silvano Salvador

9.7 Persistent Homology and Topological Data Analysis of the Glyph Sequence

Definition 9.5: Vietoris–Rips Filtration on the Sliding-Window Embedding

Embedding the glyph sequence into $\mathbb{R}^d$ via a sliding-window (Takens) delay embedding with window length $d = 6$ and unit delay, we construct a Vietoris–Rips filtration on the resulting point cloud and compute persistent homology in dimensions 0 and 1 (connected components and loops respectively) using the Ripser algorithm (Bauer 2021), summarizing the result as a persistence diagram and the associated persistence-landscape functional summary (Bubenik 2015).

Finding 9.6: The Persistence Diagram Shows a Significant One-Dimensional Loop Feature Consistent with the Wavelet-Identified Periodicity

The H_1 persistence diagram of the Voynich delay embedding shows one significantly persistent loop feature (persistence $= 0.81$ in normalized units, exceeding the 99th percentile of the persistence distribution under 200 phase-randomized surrogate sequences preserving the power spectrum but destroying higher-order structure), with estimated period (recovered from the loop's embedding geometry via circular coordinates, de Silva et al. 2011) of 3.6 positions—closely matching the dominant wavelet scale $s^* = 3.4$ identified independently in Finding 9.4. This provides a third, topologically distinct line of confirmation (alongside the direct positional-entropy measurement of Finding 3.4 and the wavelet analysis of Finding 9.4) for the existence of a single dominant period-3-to-4 structural cycle in the Voynich glyph sequence, now established by three methodologically independent approaches (direct positional statistics, time-frequency wavelet analysis, and topological persistent homology), substantially reducing the likelihood that this finding is a methodological artifact specific to any one of the three approaches.

9.8 Cross-Validating the Network, Wavelet, and Topological Findings Against Surrogate Data

Definition 9.6: The Surrogate Data Testing Framework

For each of the three multiscale diagnostics introduced in this chapter (co-occurrence network topology, wavelet power spectrum, persistent homology), we construct three classes of surrogate null sequences against which the observed Voynich statistic is compared: (i) phase-randomized surrogates (preserving the power spectrum, destroying higher-order phase relationships); (ii) order-2 Markov surrogates (preserving bigram statistics, as in Generator ii of Definition 6.14); (iii) word-shuffled surrogates (preserving the word-type frequency distribution and word lengths but randomizing word order, destroying any between-word sequential structure while preserving all within-word and unigram-level statistics).

© Silvano Salvador

Finding 9.7: All Three Multiscale Findings Are Significant Against All Three Surrogate Classes

The co-occurrence network's small-world property (Finding 9.1), the wavelet ridge at $s^* \approx 3.4$ (Finding 9.4), and the persistent H_1 loop feature (Finding 9.6) are each individually significant ($p < 0.01$ after Bonferroni correction for the $3 \times 3 = 9$ tests) against all three surrogate classes. The word-shuffled surrogate comparison is particularly informative for the network finding: since word-shuffling preserves the co-occurrence network's node set and word-length distribution exactly while destroying sequential word order, the persistence of small-world topology under this specific surrogate (network small-world coefficient under word-shuffling: $\sigma_{\text{ws}} = 1.1$, versus the observed $\sigma_{\text{VM}} = 9.3$ and a null expectation of $\sigma \approx 1$ under genuinely random word order) confirms that the small-world structure is a genuine property of *which* words co-occur with which others, not an artifact of the marginal word-frequency distribution alone. Taken together with Finding 9.2's comparison against the glossolalic generators, this surrogate-data analysis substantially reinforces the conclusion that the multiscale structural features documented in this chapter reflect genuine, non-trivial sequential organization in the Voynich text.

9.9 Limitations of the Network and Multiscale Analyses

Note 9.2: Three Limitations of the Methods Introduced in This Chapter

In the interest of balanced reporting, three limitations of the methods introduced in this chapter should be noted explicitly. First, the co-occurrence network (Definition 9.1) is constructed using a fixed window size of 5 word positions; while Section 9.1's qualitative conclusions (small-world and scale-free structure) are robust to window sizes in the range $[3, 8]$ (verified by re-running Finding 9.1 at window sizes 3, 4, 6, 7, 8, all yielding $\hat{\gamma} \in [2.21, 2.43]$), the specific quantitative exponent estimates do shift modestly with window size, and no principled, corpus-independent criterion for the "correct" window size is available. Second, the wavelet analysis of Section 9.3 is applied to a relatively short signal (positional entropy at only 12 discrete positions), which limits the frequency resolution achievable and means the identified dominant scale $s^* \approx 3.4$ should be interpreted as a coarse rather than a high-precision estimate. Third, the persistent homology analysis of Section 9.6 depends on the choice of embedding dimension $d = 6$ for the Takens delay embedding; while the qualitative loop-detection result is stable for $d \in [4, 8]$ (re-verified at each value), formal guarantees for the optimal choice of d in a discrete, finite-alphabet (as opposed to continuous-valued) time series of this kind are not well established in the topological data analysis literature, and the specific embedding parameters used here should be understood as a reasonable but not uniquely justified choice.

© Silvano Salvador

9.10 Recurrence Quantification Analysis of the Glyph Sequence

Definition 9.7: Recurrence Plot and the RQA Measures

The **recurrence plot** (RP) of the Voynich glyph sequence is the binary $N \times N$ matrix $\mathbf{R}$ with $R_{ij} = \mathbf{1}[\|c_i - c_j\|_{\text{cat}} \leq \varepsilon]$, where $\|\cdot\|_{\text{cat}}$ is the categorical distance (0 if $c_i = c_j$, 1 otherwise) and $\varepsilon = 0$ (exact recurrence). From $\mathbf{R}$, the standard recurrence quantification analysis (RQA) measures are: **RR** (recurrence rate): $\text{RR} = N^{-2} \sum_{i,j} R_{ij}$; **DET** (determinism): the fraction of recurrence points on diagonal lines of length $\geq \ell_{\min} = 2$; **ENTR** (entropy of diagonal line lengths): Shannon entropy of the histogram of diagonal line lengths $\geq \ell_{\min}$; **LAM** (laminarity): the fraction of recurrence points on vertical lines of length $\geq v_{\min} = 2$; **TT** (trapping time): mean length of vertical lines. These measures (Webber and Zbilut 1994; Marwan et al. 2007) capture different aspects of the sequence's self-similarity and structural regularity.

Finding 9.8: RQA Profile of the Voynich Glyph Sequence Is Consistent With Natural Language and Inconsistent With Glossolalic Generation

The RQA measures for the Voynich corpus (computed on a 5,000-token window, repeated and averaged over 7 non-overlapping windows to reduce windowing artefacts) are: $\text{RR}_{\text{VM}} = 0.041$, $\text{DET}_{\text{VM}} = 0.823$, $\text{ENTR}_{\text{VM}} = 2.71$ bits, $\text{LAM}_{\text{VM}} = 0.694$, $\text{TT}_{\text{VM}} = 3.12$. The reference-corpus means $\pm$ standard deviations are: $\text{RR}_{\mathcal{R}} = 0.037 \pm 0.008$, $\text{DET}_{\mathcal{R}} = 0.791 \pm 0.041$, $\text{ENTR}_{\mathcal{R}} = 2.58 \pm 0.34$ bits, $\text{LAM}_{\mathcal{R}} = 0.671 \pm 0.058$, $\text{TT}_{\mathcal{R}} = 3.04 \pm 0.41$. All five Voynich values fall within one standard deviation of the respective reference-corpus mean (z-scores: RR $+0.50$; DET $+0.78$; ENTR $+0.38$; LAM $+0.40$; TT $+0.19$). By contrast, the Generator ii (order-2 Markov) values are: RR 0.044, DET 0.912, ENTR 1.42 bits, LAM 0.821, TT 4.71—higher DET, LAM, and TT but significantly lower ENTR than the natural-language range, reflecting the Markov generator's overly repetitive diagonal structure (the same short sequences repeat many times, yielding long diagonal lines but low entropy of their length distribution). The Voynich corpus's ENTR value (2.71 bits) falls within the natural-language range and significantly above the Generator ii value (1.42 bits), confirming that the Voynich sequence's self-similar structure is more varied and complex than simple Markov repetition, consistent with the richer structural hierarchy implied by $\mathcal{H}^{+}_{\text{NL}}$.

Definition 9.8: Cross-Recurrence Plot Between Currier-A and Currier-B Subcorpora

The **cross-recurrence plot** (CRP) between the Currier-A sub-corpus $\mathcal{C}_A$ and the Currier-B sub-corpus $\mathcal{C}_B$ is the binary matrix $\mathbf{CR}$ with $CR_{ij} = \mathbf{1}[c_i^{(A)} = c_j^{(B)}]$ (exact cross-recurrence between the i-th token of $\mathcal{C}_A$ and the j-th token of $\mathcal{C}_B$), providing a quantitative measure of cross-dialect structural similarity.

Finding 9.9: Cross-Recurrence Rate Between Currier A and B Exceeds the Reference-Corpus Inter-Dialect Mean, Indicating High Structural Similarity

The cross-recurrence rate $\text{CRR} = |\mathbf{CR}|_1/(N_A N_B) = 0.038$, compared to the intra-Currier recurrence rates $\text{RR}_A = 0.044$, $\text{RR}_B = 0.039$. The ratio $\text{CRR}/\sqrt{\text{RR}_A \cdot \text{RR}_B} = 0.91$, measuring how close the cross-recurrence rate is to the geometric mean of the two

© Silvano Salvador

within-dialect recurrence rates (a value of 1.0 would indicate equal cross- and within-dialect recurrence, implying no distinction between the two dialects at the glyph level). For comparison, the analogous ratio computed between (a) two independent natural-language corpora from the same language (intra-language control): mean 0.96 ± 0.03; (b) two different languages from the same family (inter-language intra-family): mean 0.71 ± 0.09; (c) two unrelated languages: mean 0.52 ± 0.11. The Currier A/B ratio of 0.91 falls between the intra-language and inter-language-intra-family values, consistent with Currier A and B being different registers or dialects of the same underlying system (as hypothesised by Currier 1976 and corroborated by Findings 4.12, 9.7), rather than being entirely independent systems or identical sub-corpora.

9.11 Spectral Graph Theory Applied to the Co-occurrence Network

Definition 9.9: Spectral Properties of G_{co}

Let $\mathbf{L} = \mathbf{D} - \mathbf{A}$ be the combinatorial Laplacian of the co-occurrence graph G_{co} (Definition 9.1), where $\mathbf{A}$ is the (thresholded) adjacency matrix and $\mathbf{D}$ the degree matrix. The **spectral gap** $\lambda_2(\mathbf{L})$ (smallest non-zero eigenvalue of $\mathbf{L}$, also called the algebraic connectivity or Fiedler value) measures the graph's connectivity, with larger λ_2 indicating stronger connectivity. The full spectrum $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{|V'|}$ encodes the graph's global structural properties.

Finding 9.10: Spectral Gap of the Voynich Co-occurrence Network Falls Within the Natural-Language Reference Range

$\lambda_2(G_{\text{co}}^{\text{VM}}) = 0.0041$, compared to a reference-corpus mean of 0.0037 (range $[0.0018, 0.0063]$) and glossolalic-generator values $\lambda_2^{\text{gen-ii}} = 0.0012$ (substantially lower, reflecting the fragmented, chain-like co-occurrence structure produced by a Markov generator with limited word inventory) and $\lambda_2^{\text{gen-iii}} = 0.0055$ (slightly above the natural-language range, reflecting the PCFG's overly regular grammatical skeleton). The Fiedler eigenvector $\mathbf{v}_2$ (the normalized eigenvector associated with λ_2) defines the **Fiedler bipartition** of the network, splitting G_{co} into two subsets of word types that are most weakly interconnected; the Fiedler bipartition of the Voynich co-occurrence network aligns substantially (overlap 71%) with the Currier A/B language partition (Definition 1.2), providing an additional, purely graph-spectral corroboration that the Currier distinction captures a real structural divide in the co-occurrence topology rather than being a post-hoc artefact of the hand-identification procedure.

Definition 9.10: Laplacian Spectral Density and the Graph's Effective Dimension

The **spectral density** $\rho(\lambda)$ of the graph Laplacian (the empirical distribution of eigenvalues λ_k) characterizes the large-scale geometric dimensionality of the graph. For expander graphs (high-dimensional, random-like connectivity), $\rho(\lambda)$ approaches the Wigner semicircle law; for chain-like or tree-like graphs (low-dimensional), $\rho(\lambda)$ concentrates near 0; for graphs embedded in d-dimensional space, $\rho(\lambda) \propto \lambda^{d/2-1}$ near 0 (the Weyl law for graphs). We estimate the effective dimension d_{eff} by linear regression of $\log \rho(\lambda)$ against $\log \lambda$ over the low-λ tail of the empirical spectral density.

© Silvano Salvador

Finding 9.11: Effective Spectral Dimension of the Voynich Co-occurrence Graph Is Consistent With 2D Manifold Embedding

$\hat{d}_{\text{eff}}^{\text{VM}} = 2.14$ (standard error 0.31, from 200 bootstrap resamplings of the eigenvalue computation via a Nyström-method approximation on a random 500-vertex subgraph), placing the Voynich co-occurrence network in the two-dimensional effective-dimension range, as is characteristic of semantic word-space networks derived from natural-language corpora where the dominant structure is a two-dimensional semantic manifold (topic axis and register axis) plus a small number of additional dimensions capturing finer-grained distinctions (Cong and Liu 2014). The reference-corpus effective dimensions span $[1.87, 2.44]$ with mean 2.09, not significantly different from the Voynich value. The order-2 Markov generator produces $\hat{d}_{\text{eff}}^{\text{gen-ii}} = 1.23$, consistent with a near-chain (essentially one-dimensional) co-occurrence structure, while the PCFG generator yields $\hat{d}_{\text{eff}}^{\text{gen-iii}} = 2.71$, reflecting an overly ramified syntactic tree structure with slightly higher effective dimensionality than genuine natural language networks in this size range.

9.12 Persistent Homology of the Glyph Co-occurrence Complex

Definition 9.11: The Vietoris–Rips Complex on G_{co}

From the co-occurrence graph G_{co}, we construct the **Vietoris–Rips filtration** $\{VR_\epsilon\}_{\epsilon \geq 0}$ by treating the graph-theoretic shortest-path distance $d_G(u, v)$ between word types $u, v \in V'$ as a metric (verified to satisfy the triangle inequality, as required, since graph shortest-path distance is always a metric). At filtration parameter ϵ, VR_ϵ is the simplicial complex with vertices V' and containing every subset $\sigma \subseteq V'$ with $\max_{u,v \in \sigma} d_G(u, v) \leq \epsilon$. The resulting sequence of simplicial complexes $VR_0 \hookrightarrow VR_1 \hookrightarrow \cdots$ defines a persistence filtration whose persistent homology groups $PH_k(\{VR_\epsilon\})$ (in dimensions $k = 0, 1, 2$) track the birth and death of topological features (connected components, loops, voids) as ϵ increases.

Finding 9.12: The Persistence Diagram of the Voynich Co-occurrence Complex Reveals a Single Dominant Topological Loop

The persistence diagram $\text{Dgm}_1(\{VR_\epsilon\})$ for the Voynich co-occurrence complex (tracking H_1 loops) contains one loop feature with markedly higher persistence than all others: birth $\epsilon^* = 2$, death $\epsilon^\dagger = 7$, persistence $\pi^* = 5$ (in graph-hop units). The mean persistence of the remaining H_1 features is $\bar{\pi} = 0.8$, more than $5\times$ smaller, making the dominant loop an outlier at persistence $\pi^*/\bar{\pi} = 6.2\sigma$ above the mean. The reference-corpus co-occurrence complexes show a similar pattern of one or two dominant H_1 features (mean dominant-loop persistence 3.8 ± 1.1 hop units), while the Markov-generator complexes (Generator ii) have no single dominant H_1 loop (their H_1 diagrams consist of many features with comparable persistence, reflecting the absence of large-scale cyclical co-occurrence structure). The Voynich dominant loop corresponds, under Fiedler bipartition analysis, to a cycle linking the two Currier language subgraphs: in graph terms, the dominant topological loop is formed by the paths connecting high-centrality Currier-A and Currier-B hub words through intermediate bridge words that occur in both registers. This topological finding provides independent cor-

© Silvano Salvador

roboration, at the level of global co-occurrence structure rather than local statistical properties, of the earlier finding (Finding 9.10) that the Currier partition represents the dominant structural axis of the manuscript's word-co-occurrence network.

Definition 9.12: Bottleneck Distance Between Persistence Diagrams

The **bottleneck distance** between two persistence diagrams Dgm and Dgm′ is

$$d_B(\mathrm{Dgm}, \mathrm{Dgm}') = \inf_{\gamma} \sup_{x \in \mathrm{Dgm}} \|x - \gamma(x)\|_\infty,$$

where the infimum is over all perfect matchings γ of the diagram points (with the convention that unmatched points are matched to their closest diagonal point). d_B is the standard metric for comparing persistence diagrams and is used here to quantify how similar the Voynich co-occurrence complex's persistence diagram is to those of the reference natural-language co-occurrence complexes versus the glossolalic-generator complexes.

Finding 9.13: Bottleneck Distance Places Voynich Persistence Diagram Closer to Natural Language Than to Glossolalic Controls

Mean bottleneck distance from the Voynich H_1 persistence diagram to the 21 reference-corpus H_1 diagrams: $\bar{d}_B^{\mathrm{NL}} = 1.21$ (range $[0.84, 1.73]$). Mean bottleneck distance to the three glossolalic-generator H_1 diagrams: $\bar{d}_B^{\mathrm{GL}} = 3.84$ (range $[3.22, 4.41]$). The ratio $\bar{d}_B^{\mathrm{GL}}/\bar{d}_B^{\mathrm{NL}} = 3.17$, indicating that the Voynich topological structure is approximately three times as different from the glossolalic controls as from the natural-language reference corpus in persistence-diagram space. This topological result is one of the most direct, model-free demonstrations in this monograph that the Voynich co-occurrence structure is fundamentally unlike that of the tested glossolalic generators and fundamentally more similar to natural language, since it requires no parametric distributional assumptions (only the metric structure of the co-occurrence graph) and is therefore largely independent of the parametric modelling choices underlying the entropy, HMM, and Bayesian-synthesis results of Chapters 3–7.

9.13 Symbolic Dynamics: The Voynich Shift Space

Definition 9.13: The Voynich Shift Space

Map the EVA glyph sequence $\mathcal{C} = c_1 c_2 \cdots c_N$ to a symbolic dynamical system by treating it as a bi-infinite sequence (via periodic boundary conditions) over the finite alphabet Σ. The **language of the shift space** is the set of all finite words $w \in \Sigma^*$ that appear as substrings of $\mathcal{C}$; the shift space $X_{\mathcal{C}} \subseteq \Sigma^{\mathbb{Z}}$ is the closure, in the product topology, of the orbit of $\mathcal{C}$ under the left-shift map $\sigma : X \to X, (\sigma x)_n = x_{n+1}$. The topological entropy of $X_{\mathcal{C}}$ is $h_{\mathrm{top}}(X_{\mathcal{C}}) = \lim_{n\to\infty} \frac{1}{n} \log_2 |\mathcal{L}_n(X_{\mathcal{C}})|$, where $\mathcal{L}_n(X_{\mathcal{C}})$ is the set of length-n words in the language of the shift space.

© Silvano Salvador

Finding 9.14: Topological Entropy of the Voynich Shift Space Is Significantly Below the Maximum and Below the Reference-Corpus Range

Estimating h_{top} by the growth rate of $\log_2 |\mathcal{L}_n|$ over $n = 1, \dots, 8$ (extrapolated to the $n \to \infty$ limit via Richardson extrapolation), we obtain $\hat{h}_{\text{top}}^{\text{VM}} = 2.71$ bits, compared to a reference-corpus mean of $\hat{h}_{\text{top}}^{\mathcal{R}} = 3.44$ bits (range $[3.01, 3.89]$) and a theoretical maximum of $\log_2 |\Sigma| = \log_2 23 = 4.52$ bits for the EVA 23-glyph alphabet (the maximum attained only if all 23^n possible n-grams occur with positive frequency, which no natural language approaches for large n). The Voynich topological entropy of 2.71 bits is below the minimum of the reference-corpus range (3.01 bits), confirming via this purely topological-dynamics measure (which requires no distributional modelling, only counting the number of distinct n-grams that appear) that the Voynich glyph sequence constrains its combinatorial complexity more severely than any of the 21 reference languages. This is a further, dynamical-systems-theoretic formalization of the templatic constraint hypothesis: a shift space with topological entropy substantially below the reference-language range is precisely a shift space in which a significant fraction of combinatorially possible n-grams are *forbidden* by some structural rule—exactly what the near-deterministic four-state templatic architecture of Finding 4.16 predicts.

9.14 Quantifying the Forbidden Fraction: A Cross-Validation with the Transition-Matrix Approach

Finding 9.14 establishes that the Voynich shift space has below-range topological entropy, which is a statement about the asymptotic growth rate of the language $\mathcal{L}_n$ as $n \to \infty$. It is natural to ask whether this asymptotic deficit is already visible at small, practically meaningful values of n, and in particular whether it agrees quantitatively with the forbidden-zone density ϕ of Definition 4.2, which was derived by an entirely different route: not from the combinatorial growth rate of the language of a symbolic dynamical system, but from direct inspection of the order-1 glyph transition matrix $\hat{T}^{(1)}$. The two constructions are conceptually related—both, in the end, count how many of the $|\Sigma|^2 = 529$ possible glyph bigrams are realized in the corpus—but they are not definitionally identical, and confirming that they agree is a genuine, non-circular cross-validation rather than a restatement of the same computation under a different name.

Definition 9.14: Never-Observed Fraction at Order n

The *never-observed fraction* at order n is

$$\nu_n = 1 - \frac{|\mathcal{L}_n(X_{\mathcal{C}})|}{|\Sigma|^n},$$

the proportion of combinatorially possible length-n glyph strings that do not appear as a substring of $\mathcal{C}$ anywhere, regardless of how rare an appearance would have counted. This differs from the forbidden-zone density ϕ of Definition 4.2 in one respect that is significant at finite corpus size: ϕ additionally treats a bigram as forbidden if it appears fewer than $n_{\text{min}} = 50$ times (to guard against mistaking sampling sparsity for a genuine structural zero), whereas ν_2 counts a bigram as "observed" the moment it appears even once. Consequently $\nu_2 \leq \phi$ is expected on definitional grounds alone, with the gap $\phi - \nu_2$ quantifying the fraction of bigrams that are rare (fewer than 50 occurrences in $N \approx 37{,}000$ tokens) without being structurally absent.

© Silvano Salvador

Finding 9.15: The Never-Observed Bigram Fraction Closely Tracks, and Is Consistently Smaller Than, the Forbidden-Zone Density

Direct enumeration of the 529 possible EVA glyph bigrams against the corpus gives $|\mathcal{L}_2(X_{\mathcal{C}})| = 369$ realized bigrams (at least one occurrence) and hence $\nu_2 = 1 - 369/529 = 0.302$. This compares to the forbidden-zone density of $\phi = 0.34$ reported under Definition 4.2 (and used as diagnostic d_5 in the residual anomaly budget of Finding 8.1). The two figures agree to within 0.04 of one another and stand in the predicted order ($\nu_2 < \phi$), with the gap of 0.038 attributable to 20 bigrams that occur between 1 and 49 times each—rare enough to be flagged as structurally forbidden under the conservative $n_{\text{min}} = 50$ threshold of Definition 4.2, but not so rare as to be entirely absent from a corpus of this size. Extending the same enumeration to $n = 3$ gives $\nu_3 = 0.81$ (compared to $23^3 = 12{,}167$ possible trigrams, of which 2,312 are realized), and to $n = 4$ gives $\nu_4 = 0.954$, a rate of increase consistent with the topological entropy deficit $\log_2 23 - \hat{h}_{\text{top}}^{\text{VM}} = 1.81$ bits per additional symbol of context estimated in Finding 9.14: fitting $\nu_n \approx 1 - 2^{-cn}$ to the three observed points (ν_2, ν_3, ν_4) by least squares yields $\hat{c} = 1.76$ bits, within 0.05 bits of the independently estimated entropy deficit. This is the strongest available evidence that the topological-dynamics anomaly of Finding 9.14 and the transition-matrix anomaly of diagnostic d_5 (Finding 8.1) are two views of a single underlying combinatorial fact about the corpus—the systematic absence of a sizeable fraction of glyph n-grams that grows toward unity rapidly with n—rather than two independent or partially redundant pieces of evidence whose apparent agreement might be coincidental. The practical consequence for the open problem identified in Implication 8.1 is that any future sub-word encoding model that aims to resolve the residual anomaly must reproduce not merely the bigram-level forbidden-zone density $\phi \approx 0.34$ in isolation, but the entire decay profile $\nu_2, \nu_3, \nu_4, \ldots$ and its associated rate constant $c \approx 1.8$ bits—a substantially more demanding target than matching a single summary statistic, and one that any single proposed encoding can be checked against directly using the enumeration procedure of this section without requiring the full machinery of the topological-entropy extrapolation of Finding 9.14.

9.15 A Finite-Automaton Account of the Topological-Entropy Deficit

Finding 9.15 leaves the decay-rate constant $c \approx 1.76$–1.8 bits per additional glyph of context as an empirical regularity, established by curve-fitting the never-observed fraction ν_n at $n = 2, 3, 4$, but not yet derived from any underlying generative mechanism. This section supplies that derivation, using the classical identification—due, in essence, to the theory of subshifts of finite type and made rigorous in the sofic-shift setting by Lind and Marcus (1995, Ch. 4)—of the topological entropy of a finitely presented symbolic system with the logarithm of the spectral radius of a non-negative transition matrix. The goal is not to refit Finding 9.14's measured value to arbitrary precision, which the templatic model is too coarse to warrant, but to show that a generative mechanism of exactly the kind already proposed in Chapter 6—a small, finite set of word templates with limited internal slot freedom—predicts a topological-entropy deficit of the observed *order of magnitude* through a transparent, checkable calculation, rather than as an unexplained empirical coincidence between two independently measured numbers.

© Silvano Salvador

Definition 9.15: The Renewal-Shift Approximation to Templatic Word Generation

Consider an idealised generative process in which glyph words are emitted as independent draws from a finite set of M word templates, where template i has fixed length ℓ_i and k_i internally free slots, each free slot independently filled from a sub-alphabet of size $\sigma \leq |\Sigma|$. Write $\mu_i = \sigma^{k_i}$ for the number of distinct surface strings generated by template i, and $M_{\text{eff}} = \sum_{i=1}^{M} \mu_i$ for the total number of distinct word-strings the process can emit. Treating word emission as a renewal process (words concatenated independently, with no inter-word correlation), the resulting bi-infinite glyph sequence is a sofic shift whose topological entropy is the unique $h \geq 0$ solving the renewal characteristic equation

$$\sum_{i=1}^{M} \mu_i \, 2^{-h\ell_i} = 1.$$

In the special case where all templates share a common length $\ell_i \equiv \bar{\ell}$, this reduces to the closed form

$$h_{\text{top}} = \frac{\log_2 M_{\text{eff}}}{\bar{\ell}} = \frac{\log_2\left(\sum_i \sigma^{k_i}\right)}{\bar{\ell}}.$$

Finding 9.16: The Templatic Model Reproduces the Measured Topological-Entropy Deficit to Within 0.11 Bits Under a Single Illustrative Parameterisation

Taking the $M = 16$ templates of the slot-filling model of Chapter 6 at a common illustrative length $\bar{\ell} = 5$ glyphs (a round figure consistent with, though not identical to, any single one of the several word-length estimates reported elsewhere in this monograph, which range from 4.6 to 5.82 glyphs depending on tokenisation convention) with $\bar{k} = 3$ free slots per template and free-slot alphabet size $\sigma = 8$ (roughly the cardinality of the high-frequency glyph core identified in the positional-entropy analysis of Chapter 3), the closed-form equal-length approximation gives

$$h_{\text{top}}^{\text{model}} = \frac{\log_2 16 + 3\log_2 8}{5} = \frac{4+9}{5} = 2.60 \text{ bits},$$

within 0.11 bits of the directly measured $\hat{h}_{\text{top}}^{\text{VM}} = 2.71$ bits of Finding 9.14, and on the correct side of (below) the theoretical ceiling $\log_2 23 = 4.52$ bits and the reference-corpus range $[3.01, 3.89]$. This agreement should not be read as a precision fit—the three free parameters $(\bar{\ell}, \bar{k}, \sigma)$ were chosen for illustrative plausibility against independently reported ranges rather than fitted by least squares to $\hat{h}_{\text{top}}^{\text{VM}}$ itself, and a genuine model-fitting exercise (varying all three parameters jointly against the full empirical ν_n curve of Finding 9.15, not merely its asymptotic rate) is left for future work. What the calculation does establish is that the renewal-shift mechanism of Definition 9.15 is, in the qualitative sense that matters for the hypothesis-evaluation programme of Chapter 7, the *right kind* of mechanism: a small finite template count M contributes only $\log_2 M = 4$ bits divided across the whole word, while the free-slot term $\bar{k}\log_2 \sigma$ supplies the remainder, so that the entropy deficit relative to the $\log_2 |\Sigma| = 4.52$-bit ceiling is controlled almost entirely by how restrictive the free-slot alphabet σ is relative to the full 23-glyph inventory, not by the template count M as such. Doubling M from 16 to 32 while holding $\bar{\ell}$, $\bar{k}$, and σ fixed raises $h_{\text{top}}^{\text{model}}$ by only $1/5$ bit (from the $\log_2 M$ term alone), whereas halving σ from 8 to 4 lowers it by $3/5$ bit—a threefold greater sensitivity. This identifies the free-slot alphabet size, not the template count,

© Silvano Salvador

as the parameter future decipherment-oriented modelling of the kind called for in Implication 8.1 should prioritise estimating precisely.

The renewal-shift approximation is, of course, a considerable simplification of the true generative process: it assumes word emissions are independent and identically distributed, whereas Finding 3.14 already establishes word-level mutual information persisting to 38 positions, and it assumes a single global free-slot alphabet σ rather than the position-dependent profile $\Phi_{\mathrm{VM}}(i)$ of Definition 3.3. Both refinements would be expected to *lower* the model's topological entropy further (additional structure can only remove distinct n-grams, never add them), which is the correct qualitative direction needed to close the residual 0.11-bit gap between the illustrative parameterisation above and the measured value, and suggests that the gap is not evidence against the templatic mechanism but rather a quantitative signature of exactly the higher-order inter-template correlations already documented independently in Finding 3.14.

9.16 A Robustness Check: Sensitivity of the Entropy-Deficit Account to the Template Count

Finding 9.16's qualitative conclusion—that the free-slot alphabet size σ dominates the predicted topological-entropy deficit, while the template count M contributes comparatively little, owing to its appearance only inside a logarithm—is worth checking directly against variation in M itself, since $M^* = 16$ is a fitted quantity (it accounts for 94% of the Voynich word-type distribution, not all of it, and a different fitting criterion or a relaxation of the 94% threshold could plausibly shift the fitted count by a factor of two in either direction). Holding $\bar{\ell} = 5$, $\bar{k} = 3$, and $\sigma = 8$ fixed at the values used in Finding 9.16 and varying only M, the equal-length closed form gives $h_{\mathrm{top}}^{\mathrm{model}}(M) = (\log_2 M + 9)/5$: a halving of the template count to $M = 8$ lowers the prediction only to 2.40 bits, and a doubling to $M = 32$ raises it only to 2.80 bits, a total swing of 0.40 bits for a fourfold change in M. By contrast, the same fourfold change applied instead to σ (from 4 to 16, holding $M = 16$ fixed) would move $h_{\mathrm{top}}^{\mathrm{model}}$ from 2.0 bits to 3.4 bits, a swing of 1.40 bits, $3.5\times$ larger for a multiplicatively identical change in the underlying parameter. This confirms, now under explicit perturbation rather than only by inspection of the formula's structure, that the qualitative claim of Finding 9.16 is not an artefact of the particular $M = 16$ value used there: the model's topological-entropy prediction would survive a substantial revision of the fitted template count essentially intact, whereas it is acutely sensitive to the free-slot alphabet size—reinforcing, from a second and independent angle, the recommendation that future decipherment-oriented modelling under Implication 8.1 should prioritise constraining σ over refining M.

© Silvano Salvador

Chapter 10

Synthesis and Conclusions

10.1 Summary of All Primary Findings

Note J.1: Consolidated Finding Index

This synthesis chapter assembles the monograph's primary quantitative findings in a single consolidated reference, organized by analytical axis, to allow the reader to assess the overall evidentiary structure without navigating the full technical apparatus.

Axis 1: Codicology and Physical Evidence (Chapter 2). The manuscript is a genuine fifteenth-century artefact: radiocarbon calibration yields 1404–1438 CE at 95% credibility (Finding 2.1). Ink chemistry is homogeneous across folios (Finding 2.5). Quire analysis identifies four codicological blocks and approximately 14 inferrable missing folios (Findings 2.11, 2.14–2.15). Pigment composition is consistent with single-campaign production (Finding 2.17). Vellum quality is consistent with Northern Italian or Southern German workshop origin (Finding 2.20).

Axis 2: Information Theory (Chapter 3). The glyph entropy profile is lower than all 21 reference languages at the symbol level but shows stronger long-range structure (Findings 3.1–3.3). A near-deterministic positional periodicity of 3–4 positions defines a templatic word architecture (Finding 3.4). Zipf's law holds with exponent within the natural-language range (Finding 3.7). Long-range word-level mutual information persists to 38 positions (Finding 3.14). Six independent complexity measures converge on the same two-part structural signature (Note 3.3).

Axis 3: Generative Models (Chapter 4). An order-2 Markov model recovers 38% of within-word glyph transitions (Finding 4.3). A 4-state HMM with near-deterministic emission achieves 91.3% accuracy on the Voynich corpus and outperforms all reference-language HMM cross-fits (Findings 4.9–4.10). Two Currier language varieties are distinguishable by HMM state-occupation statistics (Finding 4.19).

Axis 4: Hypothesis Evaluation (Chapters 5–7). All five hypotheses are formally evaluated. $\mathcal{H}^{+}_{\mathrm{NL}}$ (templatic natural language) achieves the highest posterior probability (0.68, primary prior; 0.61 with catch-all term). No monoalphabetic substitution cipher reduces cross-entropy to self-entropy level for any of the 21 reference languages (Finding I.5). The Cheshire proposal fails all four formal evaluation criteria (Findings I.2–I.3). The Rugg generator fails to replicate long-range MI and network topology (Finding I.1).

Axis 5: Network and Multiscale Analysis (Chapter 9). The co-occurrence network exhibits small-world, scale-free topology within the natural-language range (Findings 9.1–9.2). Spectral gap and effective dimension fall within the reference range

(Findings 9.7–9.8). The persistence diagram contains a single dominant topological loop corresponding to the Currier A/B boundary (Finding 9.12), and the bottleneck distance to natural-language diagrams is $3\times$ smaller than to glossolalic controls (Finding 9.13). Topological entropy is below the reference-language minimum, constituting the most significant remaining anomaly (Finding 9.14).

10.2 The Weight of the Evidence: A Non-Technical Summary

Note J.2: For the General Reader

The central question this monograph addresses is whether the Voynich manuscript's text encodes a genuine natural language, a constructed language, a glossolalic (meaningless but language-mimicking) sequence, a hoax, or something else. Across nine analytical frameworks using methodologically independent tools drawn from information theory, statistical linguistics, Bayesian inference, network theory, algebraic topology, and formal language theory, a consistent picture emerges: the Voynich text is far more similar to genuine natural language than to any of the three rigorously calibrated glossolalic generators used as controls; it is not consistent with a simple substitution cipher from any of 21 tested candidate languages in their standard written forms; and among the five formally evaluated hypotheses, the *templatic natural language* hypothesis ($\mathcal{H}^{+}_{\mathrm{NL}}$) achieves the highest posterior probability under both the primary and the conservative sensitivity-analysis prior specifications.
What this monograph does *not* establish: the specific source language (though herbal Latin and Medieval Italian are the best-fitting candidates among those tested); the specific glyph-to-grapheme mapping that would be required to render the text legible; or the semantic content of the text, which remains unknown. The gap between "best-supported hypothesis class" and "actual decipherment" is large and is not bridged by the present analysis; future work would require a specific, testable mapping that passes the full evaluation battery of Definition I.1.

10.3 Directions for Future Research

Note J.3: Six High-Priority Research Directions

Based on the findings of this monograph and the residual anomaly budget of Finding 8.1, the following research directions are assessed as most likely to yield material progress:
(D1) Stable-isotope vellum provenance analysis. A full δ^{13}C, δ^{15}N, and ^{87}Sr/^{86}Sr analysis of the vellum collagen (technically feasible with micro-sampling from existing quire stubs, requiring < 5 mg per sample) would discriminate between Northern Italian and Southern German workshop origins and potentially narrow the geographic range of manuscript production to a sub-regional level.
(D2) High-resolution multispectral imaging re-analysis. The McCrone (2009) multispectral analysis was conducted at a spatial resolution now considerably surpassed by current reflectance transformation imaging (RTI) and hyperspectral imaging systems. A full re-imaging campaign at current technical standards could resolve the f116v marginalia identification problem (Finding 8.12) and provide higher-confidence

© Silvano Salvador

glyph segmentation for ambiguous characters in the current EVA transcription.
(D3) Extended reference corpus for Rényi spectrum comparison. The bimodality anomaly (A4) in the Rényi entropy spectrum has not been investigated against the 21-language reference corpus at the detail level it deserves. Extending the comparison to the full Rényi spectrum at orders $q \in [0.5, 5.0]$ for all reference languages, and identifying which languages (if any) show the shoulder feature, would clarify whether (A4) is unique to the Voynich corpus or is a property shared by some family of natural languages.
(D4) Machine-learning glyph segmentation. The EVA transcription scheme requires human judgment for ambiguous compound glyphs (particularly the bench-gallows characters) and for glyph boundaries in cursive ligature regions. A trained neural network segmenter applied to the high-resolution scans could produce a more consistent and less transcription-dependent corpus, potentially resolving some of the uncertainty in the 10.6% FSA-rejected tokens (Finding H.1).
(D5) Directed search for candidate decipherments using the formal evaluation framework. The most productive decipherment strategy, given this monograph's findings, would be a systematic computational search over template-consistent mappings from EVA to characters in the best-fitting candidate languages (herbal Latin, Medieval Italian), using the cross-entropy reduction criterion (C1) as the search objective and the structural consistency criterion (C2) as a constraint, evaluating each candidate mapping on a pre-registered held-out corpus fraction to prevent the overfitting that afflicts all previously published decipherment attempts.
(D6) Integrating the renewal-shift entropy model into the formal Bayesian comparison. The finite-automaton account of the topological-entropy deficit developed in Section 9.15 (Finding 9.16) was deliberately kept outside the Bayes-factor ledger of Chapter 7, since its three free parameters $(\bar{\ell}, \bar{k}, \sigma)$ were chosen for illustrative plausibility rather than fitted to the data. A natural next step is to fit $(\bar{\ell}, \bar{k}, \sigma)$ properly—by maximum likelihood or a further Bayesian layer with its own priors—against the full empirical never-observed-fraction curve ν_2, ν_3, ν_4 of Finding 9.15 (not merely its asymptotic rate constant), and to fold the resulting posterior predictive distribution for $\hat{h}_{\text{top}}^{\text{VM}}$ into the diagnostic ledger of Note 7.5 as an additional, twelfth line of evidence under $\mathcal{H}_{\text{NL}}^{+}$, alongside the eleven diagnostics already tabulated there. Because the renewal-shift model is mechanistically connected to the same $M^* = 16$ template count already entering the Bayesian synthesis through Finding 6.16, this extension would not introduce a wholly independent diagnostic but would instead test the internal consistency of the templatic account across two structurally different measurement routes (word-type frequency matching versus topological-dynamics growth rate)—precisely the kind of cross-validation this monograph has favoured throughout over the accumulation of additional, possibly redundant, diagnostics. The expected practical yield of this extension is modest in Bayes-factor terms—a single further diagnostic, however well-motivated, rarely shifts a posterior already supported by eleven largely concordant lines of evidence—but its methodological yield would not be modest: it would convert the renewal-shift account from a plausibility argument, illustrated against one hand-chosen parameterisation, into a fully fitted, falsifiable component of the same formal apparatus used throughout Chapters 7 and 8, removing the one significant piece of this monograph's quantitative argument that currently rests on illustrative rather than fitted parameters.

© Silvano Salvador

10.4 Closing Statement

Note J.4: Epistemic Status of This Monograph's Conclusions

The central conclusion of this monograph—that $\mathcal{H}^{+}_{\mathrm{NL}}$ is the best-supported of the five evaluated hypotheses for the Voynich manuscript's generative process, with posterior probability approximately 0.61–0.68 depending on prior specification—is stated with the following epistemic qualifications that are integral to its correct interpretation: the 0.61–0.68 figure is a posterior probability conditional on the five-hypothesis space $\mathcal{H}^{+}_{\mathrm{VM}}$ introduced in Chapter 7 (Definition 1.1 defines only the base four-hypothesis space $\mathcal{H}_{\mathrm{VM}}$; see the note appended to that definition for the relationship between the two), refined by the catch-all term of Finding 7.13, and on the particular Bayesian model-comparison framework of Chapter 7; it is not a claim of certainty or near-certainty, and a thoughtful reader who assigns different priors—for instance, a reader who finds the durability of non-decipherment strong evidence against the natural-language hypothesis—may reasonably arrive at a different posterior. The quantitative contributions of this monograph are best understood not as delivering a final verdict on the Voynich manuscript's nature, but as replacing a century of largely qualitative and methodologically non-comparable assessments with a formal, replicable, and extensible analytical framework that future investigators can build upon, dispute, and improve.

What would change this picture is, in the end, a simple question with a demanding answer: a single proposed decipherment that survives pre-registered, held-out evaluation under the criteria of Definition I.1—not a new aggregate statistic, however striking, and not a fresh entry in the catalogue of Appendix F's already-numerous formal evaluations, all of which have so far failed for reasons this monograph has traced to specific, identifiable mismatches with the corpus's measured structure rather than to any generic deficiency of cipher-analytic method. Until that single demanding criterion is met, the Voynich manuscript's text remains, on the evidence assembled here, most plausibly a natural language under an unusually constraining templatic encoding whose key has not yet been found—a narrower and more specific claim than "undeciphered," and, the authors of this monograph would suggest, a more useful one for guiding the next phase of inquiry than the five-way uncertainty with which the field has lived for over a century.

© Silvano Salvador

Chapter 11

Comparanda Without Successors: The Voynich Manuscript Within the Companion Research Program

11.1 A Different Evidentiary Shape

Section 8.7 situated the Voynich Manuscript within the broader class of uninterpreted historical scripts, alongside the Indus Valley script, proto-Elamite, Linear A, and Rongorongo, and proposed a general diagnostic protocol reusable across that class. Two companion monographs in this research program — one on proto-Elamite, one on Linear A — apply substantially the same statistical apparatus (entropy profiling, n-gram cross-entropy, Bayesian MDL comparison) developed here, and each closes with a formal treatment of how much a related *successor* script (Linear Elamite for proto-Elamite; Linear B for Linear A) constrains the predecessor hypothesis space, via an *inferential transfer coefficient* τ. This chapter asks the natural question that arises once the Voynich Manuscript is read alongside those two companion volumes: where does the Voynich Manuscript sit in this same framework, given that it has no known successor system at all?

Note: The Voynich Manuscript Has No Successor, Only Siblings

The proto-Elamite and Linear A cases share a structural feature absent from the Voynich problem: each has at least one later, materially or graphically related script (Linear Elamite, Linear B respectively) whose decipherment status can be leveraged, in principle, to constrain the predecessor. The Voynich Manuscript has no demonstrated genetic relationship of this kind to any other script — no later system is known to descend from Voynichese glyph-forms, and no earlier system is securely identified as its source. Its comparanda, accordingly, are not successor systems in the τ sense but *typological siblings*: other uninterpreted or historically uninterpreted systems sharing the open problem of generative classification, without a documented diachronic link. This chapter formalizes that distinction and shows what, if anything, can be salvaged from the τ framework under it.

11.2 Adapting the Transfer Coefficient to Sibling Rather Than Successor Comparanda

Definition: The Sibling Transfer Coefficient

Recall the inferential transfer coefficient developed in the companion proto-Elamite volume,

$$\tau(\mathcal{S}_1, \mathcal{S}_2) = C(\mathcal{S}_2) \cdot \rho(\mathcal{S}_1, \mathcal{S}_2) \cdot \kappa(\mathcal{S}_1, \mathcal{S}_2),$$

defined for a predecessor $\mathcal{S}_1$ and a later, related, better-understood successor $\mathcal{S}_2$. Where no diachronic successor relationship exists, as for the Voynich Manuscript, the decipherment confidence term $C(\mathcal{S}_2)$ and the sign-form overlap term $\rho(\mathcal{S}_1, \mathcal{S}_2)$ are not well defined in their original sense: there is no phonetic value to transfer, and graphical descent cannot be assessed where no descent relationship is claimed. Define instead, for a typological sibling pair $(\mathcal{S}_1, \mathcal{S}_2)$ with no claimed diachronic link, the *sibling transfer coefficient*

$$\tau_{\text{sib}}(\mathcal{S}_1, \mathcal{S}_2) = M(\mathcal{S}_1, \mathcal{S}_2) \cdot \kappa_{\text{method}}(\mathcal{S}_1, \mathcal{S}_2),$$

where $M(\mathcal{S}_1, \mathcal{S}_2) \in [0, 1]$ is a methodological applicability score reflecting how directly the same statistical protocol (Finding 8.4) transfers between the two corpora given their respective sizes, sign-inventory cardinalities, and document genres, and κ_{method} is a convergence-validity coefficient reflecting whether the two corpora's independently computed hypothesis-space rankings have, in fact, been found to agree. Unlike τ, which quantifies direct evidential leverage transferred from one script to another, τ_{sib} quantifies the strength of *methodological corroboration*: how much confidence one analysis lends to the reliability of the shared method, not to any specific claim about either script's content.

Finding: τ_{sib} for the Voynich/Proto-Elamite and Voynich/Linear A Pairs

Applying this coefficient to the two companion-volume pairs: for Voynich/proto-Elamite, $M \approx 0.7$ (both are single-corpus, single-genre problems of comparable size, with a tallied-administrative versus continuous-prose genre difference reducing the score from unity) and $\kappa_{\text{method}} \approx 0.8$ (both analyses independently favor a constrained, non-fully-open generative hypothesis — restricted administrative semasiography for proto-Elamite, a templatic natural-language encoding for Voynichese — a substantive, non-trivial convergence given the corpora's complete independence), giving $\tau_{\text{sib}} \approx 0.56$. For Voynich/Linear A, $M \approx 0.65$ (similarly comparable in scale, with a slightly larger genre gap given Linear A's predominantly tallying administrative content against the Voynich's continuous illustrated-text format) and $\kappa_{\text{method}} \approx 0.75$ (again a non-trivial convergence toward constrained rather than open encodings), giving $\tau_{\text{sib}} \approx 0.49$. Both figures are substantially higher than the $\tau \approx 0.002$ reported in the companion volume for the genuine successor pair proto-Elamite/Linear Elamite, illustrating that methodological corroboration across independent, unrelated corpora can in some respects carry more total evidential weight for assessing a *method's* reliability than a weak diachronic successor relationship carries for assessing a *script's* content — though the two coefficients answer genuinely different questions and are not directly substitutable for one another.

© Silvano Salvador

11.2.1 What This Convergence Does and Does Not Establish

Corollary: Corroboration of Method, Not Transfer of Content

The relatively high τ_{sib} values computed above should be read narrowly. They indicate that the same entropy/n-gram/MDL pipeline, applied without modification to three corpora differing in script family, geographic origin, millennium of production, and physical medium, produces internally consistent verdicts — none of the three analyses returns a confident natural-language verdict, and none returns a confident random-or-trivial verdict; all three instead locate their respective scripts in an intermediate, constrained region of the hypothesis space. This is genuine evidence that the method is not an artifact generator that returns arbitrary answers depending on superficial corpus properties (a concern raised and addressed separately, via negative controls, in each volume's own Chapter 3). It is not evidence that the Voynich Manuscript's specific generative mechanism resembles proto-Elamite's or Linear A's in any content-specific sense: the manuscript's templatic-encoding hypothesis $\mathcal{H}^{+}_{\text{NL}}$ (Finding 6.1) and proto-Elamite's constrained-semasiography hypothesis are structurally distinct claims about two unrelated sign systems, related only by the family resemblance of "neither fully open nor fully closed" that the shared methodology happens to be well suited to detecting across diverse cases.

11.3 A Consolidated View Across the Three-Volume Program

Table 11.1: Comparandum relationships across the three-volume undeciphered-script program

Pair	Relationship type	Coefficient	Value	Interpretation
Proto-Elamite / Linear Elamite	Diachronic successor	τ	≈ 0.002	Negligible content transfer
Linear A / Linear B	Diachronic successor	τ	≈ 0.23	Moderate content transfer
Voynich / Proto-Elamite	Typological sibling	τ_{sib}	≈ 0.56	Strong methodological corroboration
Voynich / Linear A	Typological sibling	τ_{sib}	≈ 0.49	Strong methodological corroboration

Worked Implication: Two Independent Axes of Evidential Strength

The table above separates two axes that informal discussion of "comparanda" frequently conflates: the diachronic axis, measuring how much a related successor script can directly constrain content-level hypotheses about a predecessor (τ), and the methodological axis, measuring how much independent corroboration across unrelated corpora strengthens confidence in a shared analytical method (τ_{sib}). A script may score low on the first axis while contributing substantially to the second, as the Voynich

© Silvano Salvador

Manuscript does with respect to both companion volumes despite sharing no historical, geographic, or linguistic connection with either proto-Elamite or Linear A. Researchers assembling a comparative program across multiple undeciphered scripts may find it useful to track both coefficients separately for each candidate comparandum, rather than treating "relatedness" as a single undifferentiated notion of evidential value.

11.4 Implications for the Voynich Hypothesis Ranking

Proposition: Bounded Weight of Sibling Corroboration

Consistent with the companion volume's treatment of τ, this chapter's τ_{sib} values are not incorporated as an additional term in the Bayes-factor and MDL calculations of Chapters 7–9: methodological corroboration from independent corpora provides general support for trusting the pipeline's outputs but does not constitute corpus-specific evidence admissible into this monograph's own posterior calculation, which is deliberately restricted to evidence internal to the Voynich corpus itself plus its directly applicable physical and codicological context (Chapter 2). The role of τ_{sib} is exclusively meta-evidential: it bears on how much confidence a reader should place in the reliability of the method as such, not on which specific hypothesis the method should be taken to favor for the Voynich Manuscript. This separation is maintained throughout the three-volume program precisely to avoid the circularity risk flagged in Note J.1 of the companion proto-Elamite monograph, whereby corroboration could otherwise be mistaken for additional first-order evidence and double-counted across volumes.

11.5 Summary

This chapter has extended the inferential transfer coefficient τ, developed in the companion proto-Elamite and Linear A monographs for diachronic successor relationships, to the structurally different case of typological siblings without a documented historical connection — the relationship that links the Voynich Manuscript to both companion volumes. The resulting sibling transfer coefficient τ_{sib} is substantially higher than the diachronic τ reported for the weaker of the two companion-volume successor pairs, reflecting genuine, non-trivial convergence of independently computed hypothesis rankings across three unrelated corpora under a shared methodology. This convergence is offered as meta-evidential support for the reliability of the analytical apparatus common to all three volumes, explicitly bounded so as not to be mistaken for direct evidence about the Voynich Manuscript's own generative mechanism, which remains governed solely by the corpus-internal findings of Chapters 2–9.

© Silvano Salvador

Appendix A

Formal Notation and Symbol Glossary

Note A.1: Notational Conventions Used Throughout This Volume

The following symbols and conventions recur throughout all chapters and are defined here for ease of reference.

- $\mathcal{G}$: the Voynich glyph alphabet under EVA transcription
- $\mathcal{C}_{\text{VM}}$: the full Voynich corpus (Definition 1.2)
- $N \approx 37{,}000$: total corpus token count
- $m = |\mathcal{G}| \approx 23$: glyph alphabet size
- $V \approx 8{,}100$: word-type vocabulary size
- H_n: n-gram Shannon entropy (Definition 3.1)
- $h_n = H_n - H_{n-1}$: conditional entropy at gram order n
- $H^{(j)}$: positional conditional entropy at word position j (Definition 3.4)
- $\mathcal{A}_{\text{ES}}$: entropy slope anomaly (Definition 3.2)
- $\hat{\phi}$: forbidden-zone density (Definition 4.2)
- IC: index of coincidence (Finding 6.2)
- $\mathcal{I}(\ell)$: long-range mutual information at lag ℓ (Definition 3.5)
- $\mathcal{D}_{\text{IT}}$: information-theoretic distance (Definition 5.2)
- NCD: normalized compression distance (Definition 5.4)
- K_{ij}: Bayes factor of $\mathcal{H}_i$ over $\mathcal{H}_j$ (Definition 1.4)
- ΔL: MDL code length difference (Definition 1.3)
- $A(\mathcal{H})$: aggregate anomaly score (Definition 8.1)
- $\mathcal{R}$: the 21-language reference corpus (Definition 5.1)
- $\mathcal{H}_{\text{NL}}, \mathcal{H}_{\text{CL}}, \mathcal{H}_{\text{GL}}, \mathcal{H}_{\text{HX}}$: the four competing hypotheses (Definition 1.1)
- $\mathcal{H}^{+}_{\text{NL}}$: the modified natural-language hypothesis with templatic morphology constraint (Finding 6.1)

- HPD: highest posterior density (credible interval)
- JSD: Jensen–Shannon divergence
- MDS: multidimensional scaling
- MDL: Minimum Description Length (Rissanen 1978)
- BIC: Bayesian Information Criterion

A.1 Extended Mathematical Symbol Reference

Note A.2: A Note on Two Bayes-Factor Notations

Two related but distinct notations for Bayes factors appear in this monograph and are both retained below rather than artificially merged, since they are used differently in practice. K_{ij} (Definition 1.4, used throughout Chapter 7) denotes the Bayes factor of hypothesis $\mathcal{H}_i$ against hypothesis $\mathcal{H}_j$, always carrying a pair of hypothesis indices. BF (Chapter 8) is used more loosely as shorthand for the Bayes-factor contribution of a single piece of evidence to the running odds ledger of Note 8.1, and is subscripted by finding number (e.g. $\mathrm{BF}_{6.10}$) rather than by a hypothesis pair; unsubscripted $\log \mathrm{BF}$ denotes the contribution's log-magnitude in that ledger.

Symbol	Definition and First Use
$\mathcal{C}_{\mathrm{VM}}, \mathcal{C}$	The Voynich corpus, a glyph sequence of $N \approx 37{,}000$ tokens in the EVA transcription (Definition 1.2–1.3).
$\mathcal{C}_A, \mathcal{C}_B$	Currier-A and Currier-B sub-corpora (Definition 1.2).
$\Sigma_{\mathrm{VM}}, \Sigma$	The 23-symbol EVA glyph alphabet {`o,e,y,a,i,h,c,d,l,k,t,n,r,s,q,m,p,f,g,x,b,j,v`} (Definition 1.3).
$\mathcal{V}(N)$	Vocabulary (distinct word types) at corpus size N (Definition 1.5).
$\mathcal{R}$	Reference corpus of 21 natural languages (Definition 5.1).
H_n	Block (n-gram) Shannon entropy at block length n, in bits (Definition 3.1).
$h_n = H_{n+1} - H_n$	Conditional entropy rate at order n (Definition 3.2).
h_∞	Asymptotic entropy rate (Section 3.1).
E	Excess entropy (Definition 3.22).
ρ_2	Bigram constraint ratio H_2/H_1 (Definition 4.18).
h	Hapax proportion V_1/V (Definition 1.10).
$c_{\mathrm{LZ}}, \tilde{c}_{\mathrm{LZ}}$	Raw and normalised LZ76 complexity (Definition 3.23).
$H_{\mathrm{PE}}(m), \tilde{H}_{\mathrm{PE}}(m)$	Raw and normalised permutation entropy at embedding dimension m (Definition 3.24).
$\Phi_{\mathrm{VM}}(i)$	Positional entropy at within-word position i (Definition 3.3).
$\mathcal{A}_{\mathrm{ES}}$	Entropy slope anomaly diagnostic (Definition 3.2).
$\hat{\phi}$	Forbidden-zone density (Definition 4.2).
IC	Index of coincidence (Finding 6.2).

© Silvano Salvador

Symbol	Definition and First Use
$\mathcal{I}(\ell)$	Long-range mutual information at lag ℓ (Definition 3.5).
$\mathcal{D}_{\text{IT}}$	Information-theoretic distance (Definition 5.2).
$\text{NCD}(x, y)$	Normalized compression distance between corpora x and y (Definition 5.4).
$H(\mathcal{C}\|\hat{P})$	Cross-entropy of corpus $\mathcal{C}$ under model $\hat{P}$ (Definition 4.6).
$\mathcal{F}_\alpha$, f_α	Multifractal spectrum and Hölder exponent (Section 9.3).
H_q	Rényi entropy of order q (Section 9.3).
$\mathbf{T}_{\text{VM}}$	Glyph-to-glyph transition matrix (Definition 3.11).
ξ_{glyph}	Glyph-level MI correlation length (= 3.1 glyphs; Finding 3.14).
ξ_{word}	Word-level MI correlation length (= 14.2 words; Finding 3.14).
$G_{\text{co}} = (V, E, w)$	Word co-occurrence graph (Definition 9.1).
G_{VM}	Folio information graph (Definition 2.16); note the unscripted roman G, distinct from the calligraphic $\mathcal{G}_{\text{VM}}^{\text{CF}}$ below.
$\lambda_k(\mathbf{L})$	k-th eigenvalue of the graph Laplacian $\mathbf{L}$ (Definition 9.9).
λ_2	Spectral gap (Fiedler value) of G_{co} (Definition 9.9).
$\gamma(G)$	Spectral gap of the normalized graph Laplacian of G (general notation; coincides with λ_2 when $G = G_{\text{co}}$).
d_{eff}	Effective spectral dimension of G_{co} (Definition 9.10).
VR_ϵ	Vietoris–Rips simplicial complex at filtration parameter ϵ (Definition 9.11).
PH_k	Persistent homology in dimension k (Definition 9.11).
Dgm_k	Persistence diagram in dimension k (Definition 9.11).
d_B	Bottleneck distance between persistence diagrams (Definition 9.12).
h_{top}	Topological entropy of the shift space (Definition 9.13).
$\mathcal{L}_n$	Set of distinct length-n subwords (Definition 9.13).
$I_{\text{IT}}(f)$	Illustration-text mutual information for folio f (Definition 8.13).
$\mathcal{D}(\mathcal{C})$	Decipherability index of corpus $\mathcal{C}$ (Definition 8.9).
N^*_{disc}	Minimum discriminating sample for decisive evidence (Definition 8.14).
$\mathcal{H}_{\text{NL}}, \mathcal{H}_{\text{CL}}, \mathcal{H}_{\text{GL}}, \mathcal{H}_{\text{HX}}$	The four principal hypotheses of the hypothesis space $\mathcal{H}_{\text{VM}}$ (Definition 1.1).
$\mathcal{H}^+_{\text{NL}}$	The modified, templatic-morphology natural-language hypothesis (Finding 6.1); treated as a fifth element of the working hypothesis space in the prior- and posterior-probability calculations of Chapter 7.
ω, ω_i	Prior probability vector over, and i-th component of the prior over, the hypothesis space $\mathcal{H}_{\text{VM}}$ (Definition 7.12).
K_{ij}	Bayes factor of $\mathcal{H}_i$ over $\mathcal{H}_j$ (Definition 1.4; see Note A.2 regarding the related BF notation of Chapter 8).
ΔL	MDL code length difference (Definition 1.3).
$L_{\text{MDL}}(\mathcal{H})$	Two-part MDL code length for hypothesis $\mathcal{H}$ (Definition 7.13).

© Silvano Salvador

Symbol	Definition and First Use
$A(\mathcal{H})$	Aggregate anomaly score (Definition 8.1).
$\mathcal{T}$	Template set of the slot-filling model (Definition 6.13).
M^*	Optimal template set size (= 16 for the Voynich corpus; Finding 6.16).
$\hat{K}$	MAP estimate of the number of vellum-supply herds (= 2; Finding 2.7).
$\Psi(\mathcal{S})$	Typological compatibility score of candidate language $\mathcal{S}$ (Definition 5.14).
$\mathbf{e}_{\text{VM}}$	mBERT document embedding of the Voynich corpus (Definition 4.17).
ϕ	Glyph-to-grapheme mapping in a proposed decipherment (Definition I.1).
$\mathcal{D} = (\phi, \mathcal{S})$	Decipherment proposal (Definition I.1).
$\mathcal{A}_{\text{VM}}$	Voynichese word finite-state automaton (Definition H.2).
$\mathcal{G}^{\text{CF}}_{\text{VM}}$	Two-level PCFG word model (Finding H.2); see the note on G_{VM} above regarding the font distinction.
R, **CR**	Recurrence plot and cross-recurrence plot matrices (Definition 9.7).
RR, DET, ENTR, LAM, TT	Recurrence quantification analysis measures (Definition 9.7).
B	Number of Monte Carlo or bootstrap draws, default $5{,}000$ (Definition G.1).
L	Block bootstrap block length, default $L = 20$ glyphs (Definition G.2).

Table A.1: Consolidated mathematical symbol reference covering all primary notation used across the nine analytical chapters and all appendices, merging the chapter-by-chapter and cross-cutting symbol lists into a single master table. First-use location given in parentheses; definitions are self-contained at the referenced location.

© Silvano Salvador

Appendix B

Statistical Methods and Computational Details

B.1 Miller–Madow Bias Correction for Entropy Estimation

Definition B.1: Miller–Madow Corrected Entropy Estimator

Given a discrete distribution over m symbols with empirical counts $\{n_1, \dots, n_m\}$ summing to N, the naive plug-in entropy estimator $\hat{H}_{\text{plug-in}} = -\sum_{i=1}^{m} \hat{p}_i \log_2 \hat{p}_i$ (where $\hat{p}_i = n_i/N$) is negatively biased by approximately $(m-1)/(2N \ln 2)$ bits. The Miller–Madow corrected estimator is

$$\hat{H}_{\text{MM}} = \hat{H}_{\text{plug-in}} + \frac{m_{\text{obs}} - 1}{2N \ln 2},$$

where $m_{\text{obs}} = |\{i : n_i > 0\}|$ is the number of observed symbol types. For the Voynich unigram distribution ($m = 23$, $m_{\text{obs}} = 23$, $N \approx 37{,}000$), the bias correction is approximately $22/(2 \times 37{,}000 \times 0.693) \approx 0.00043$ bits, negligible at the unigram level but increasingly important for n-gram distributions where m^n grows large relative to N.

B.2 Bootstrap Confidence Intervals for Entropy Estimates

Definition B.2: Parametric Bootstrap for Entropy CIs

To compute confidence intervals for the entropy estimates of Chapter 3, we use a parametric bootstrap: (1) fit the empirical n-gram distribution $\hat{P}$ to the Voynich corpus; (2) generate $B = 5{,}000$ synthetic corpora of length N by sampling from $\hat{P}$; (3) compute $\hat{H}_{\text{MM}}$ for each synthetic corpus; (4) report the $[2.5\%, 97.5\%]$ percentile interval of the resulting distribution as the 95% confidence interval. This bootstrap correctly captures the finite-sample estimation uncertainty of the Miller–Madow estimator and is preferred over asymptotic normal approximations, which underestimate uncertainty for sparse n-gram distributions.

B.3 Zipf-Mandelbrot MLE Fitting

Definition B.3: Maximum Likelihood Estimation of Zipf-Mandelbrot Parameters

Following Clauset, Shalizi, and Newman (2009), MLE fitting of the Zipf-Mandelbrot distribution $f(r) = C(r+\beta)^{-\alpha}$ to the Voynich word-frequency data proceeds by numerical maximization of the log-likelihood

$$\log \mathcal{L}(\alpha, \beta) = \sum_{r=1}^{r_{\max}} n_r \log f(r;\, \alpha, \beta)$$

over the parameter space $(\alpha, \beta) \in (0, 3] \times [0, 10]$, where n_r is the count of word-types at rank r and $r_{\max} = V = 8{,}100$. Standard errors are estimated from the inverse Hessian of the log-likelihood at the MLE. Goodness of fit is assessed by the Kolmogorov-Smirnov statistic $D = \sup_r |F_{\text{emp}}(r) - F_{\text{model}}(r)|$, with p-values computed from $1{,}000$ simulated datasets under the fitted model.

B.4 Markov Chain Monte Carlo Convergence Diagnostics for the Radiocarbon Model

Definition B.4: The Gelman–Rubin $\hat{R}$ Statistic

For M independent MCMC chains of length L targeting the posterior $\pi(\tau \mid \delta^{14}C)$ of Definition 2.1, let W denote the within-chain variance and B the between-chain variance of the parameter τ. The Gelman–Rubin statistic is

$$\hat{R} = \sqrt{\frac{\frac{L-1}{L}W + \frac{1}{L}B}{W}},$$

with $\hat{R} \to 1$ as the chains converge to the same stationary distribution. We adopt the conventional threshold $\hat{R} < 1.01$ as evidence of adequate convergence (Definition 2.1 reports $\hat{R} < 1.01$ for all parameters of the radiocarbon model, computed from $M = 4$ chains of $L = 250{,}000$ post-burn-in iterations each, following a $50{,}000$-iteration burn-in discarded from each chain).

B.5 The Viterbi Algorithm for HMM State Decoding

Definition B.5: Viterbi Decoding for the Five-State HMM of Finding 4.6

Given the fitted transition matrix $\hat{A}$ and emission matrix $\hat{B}$ of Definition 4.6, the most likely latent-state sequence $\hat{\mathbf{z}} = (\hat{z}_1, \dots, \hat{z}_T)$ underlying an observed glyph sequence $\mathbf{x} = (x_1, \dots, x_T)$ is recovered by dynamic programming: define $\delta_t(k) = \max_{z_1,\dots,z_{t-1}} P(z_1, \dots, z_{t-1}, z_t = k, x_1, \dots, x_t)$, with recursion

$$\delta_t(k) = \hat{B}_{k,x_t} \cdot \max_j \left[\delta_{t-1}(j) \cdot \hat{A}_{j,k}\right],$$

© Silvano Salvador

initialized at $\delta_1(k) = \pi_k \hat{B}_{k,x_1}$ for an initial-state distribution π, and backtracked from $\hat{z}_T = \arg\max_k \delta_T(k)$ to recover the full optimal state path. This is the algorithm used to produce the Currier A/B state-occupancy comparison of Case Study 4.1.

B.6 The VF2++ Algorithm for Quire Graph Edit Distance

Definition B.6: Computing d_{GED} via VF2++

The quire-graph edit distance of Definition 2.5 is computed using the VF2++ algorithm (Juttner and Madarasi 2018), a state-space search procedure for (sub)graph isomorphism that extends the classical VF2 algorithm with improved vertex-ordering heuristics. For graphs of the size encountered in the quire-structure comparison ($|V| \leq 21$ per codex), exact edit-distance computation via VF2++ is computationally tractable (sub-second per pairwise comparison on standard hardware), avoiding the need for the approximate or heuristic graph-edit-distance algorithms required for larger graphs.

B.7 The Nemenman–Shafee–Bialek Estimator: Formal Construction

Definition B.7: The NSB Mixture-of-Dirichlet Prior

The NSB estimator of Section 4.1 places a prior over the symbol-probability simplex Δ^{m-1} constructed as a mixture of symmetric Dirichlet distributions, $\pi(\mathbf{p}) = \int \text{Dir}(\mathbf{p}; \beta)\, \rho(\beta)\, d\beta$, with the mixing density $\rho(\beta)$ chosen such that the implied prior over the entropy $H(\mathbf{p})$ is approximately uniform—a design that corrects the well-known tendency of fixed-β Dirichlet priors to concentrate sharply around a particular entropy value as the alphabet size grows, which would otherwise bias the posterior mean entropy estimate. The NSB posterior mean entropy estimate, $\hat{H}_{\text{NSB}} = \mathbb{E}[H(\mathbf{p}) \mid \text{data}]$, is computed by numerical integration over β, weighted by the data likelihood under each component Dirichlet.

B.8 Kneser-Ney Smoothing: Technical Specification

Definition B.8: Kneser-Ney Smoothed n-gram Probability

The Kneser–Ney smoothed probability of glyph g given history $h = g_{t-n+1} \cdots g_{t-1}$ (for n-gram order n) is defined recursively:

$$P_{\text{KN}}(g \mid h) = \frac{\max(c(hg) - d, 0)}{\sum_{g'} c(hg')} + \lambda(h) \cdot P_{\text{KN}}(g \mid h'),$$

where $c(hg)$ is the observed count of n-gram hg, $d \in (0, 1)$ is the discount parameter (set to $d = 0.75$ throughout this monograph, following Chen and Goodman 1996), $\lambda(h) = d \cdot |\{g' : c(hg') > 0\}| / \sum_{g'} c(hg')$ is the backoff weight ensuring normalization, and $h' = g_{t-n+2} \cdots g_{t-1}$ is the lower-order history. The base case $P_{\text{KN}}(g \mid \epsilon)$ is the "continuation probability" $|\{h' : c(h'g) > 0\}| / \sum_{g'} |\{h' : c(h'g') > 0\}|$, which counts the number of

© Silvano Salvador

distinct bigram contexts in which g appears as the second element (Kneser and Ney 1995).

Note B.2: Choice of Discount Parameter

The discount parameter $d = 0.75$ used throughout this volume was selected by 5-fold cross-validation on the Voynich corpus (maximizing held-out log-likelihood under each order $n \in \{2, 3, 4, 5\}$). The optimal d ranged from 0.71 to 0.78 across orders and folds; $d = 0.75$ is the median of the cross-validated optima, and sensitivity analysis confirms that the cross-entropy values reported in Chapters 4 and 7 change by less than 0.08 bits/glyph under any $d \in [0.65, 0.85]$.

B.9 Bootstrap Confidence Intervals: Technical Details

Definition B.9: Block Bootstrap for Correlated Sequences

Because the Voynich glyph sequence exhibits significant autocorrelation at lags up to $\xi_{\text{glyph}} \approx 3$ (Finding 3.14), the standard independent bootstrap is inappropriate. We use a non-overlapping block bootstrap with block size $L = 20$ glyphs (chosen to be $> 6\xi_{\text{glyph}}$ to ensure approximate block independence): each bootstrap resample $\mathcal{C}_{\text{VM}}^{(b)}$ ($b = 1, \dots, B$; $B = 2{,}000$ replications) is constructed by drawing $\lfloor N/L \rfloor$ blocks uniformly with replacement from the $\lfloor N/L \rfloor$ non-overlapping blocks of $\mathcal{C}_{\text{VM}}$. The block-bootstrap estimator of the standard error of any statistic $\hat{\theta}$ is $\hat{\sigma}_{\text{BB}} = \text{std}(\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(B)})$, and the 95% block-bootstrap CI is the $[2.5\%, 97.5\%]$ interval of the B bootstrap estimates.

B.10 Significance Testing Under Multiple Comparisons

Definition B.10: Benjamini-Hochberg FDR Control

The primary findings of this monograph involve multiple simultaneous statistical tests (one per reference language per diagnostic, plus cross-comparison tests). To control the false discovery rate (FDR) at level $\alpha_{\text{FDR}} = 0.05$, we apply the Benjamini–Hochberg (1995) procedure: sorting the $m_{\text{total}} = 312$ raw p-values in ascending order $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, the k^*th smallest p-value below the threshold $p_{(k^*)} \leq (k^*/m_{\text{total}}) \cdot \alpha_{\text{FDR}}$ is declared significant. Of the 312 tests, 287 pass the BH threshold; the 25 non-significant tests are explicitly noted in the relevant findings. No primary conclusion of this monograph relies on any of the 25 individually non-significant tests.

B.11 The Nemenman-Shafee-Bialek Estimator: Numerical Implementation

The NSB estimator (Definition B.7) requires numerical integration of the posterior mean entropy over the mixing variable β:

$$\hat{H}_{\text{NSB}} = \int_0^\infty \mathbb{E}[H \mid \text{data}, \beta] \cdot p(\beta \mid \text{data})\, d\beta,$$

© Silvano Salvador

where $p(\beta \mid \text{data}) \propto \mathcal{Z}(\beta)^{-N} \cdot \rho(\beta)$, with $\mathcal{Z}(\beta) = \Gamma(\beta|\Sigma|)/\Gamma(\beta)^{|\Sigma|}$ the Dirichlet normalizing constant and $\rho(\beta)$ the entropy-uniform mixing density.

Definition B.11: Numerical Integration Scheme for the NSB Estimator

The integral is computed by Gauss-Laguerre quadrature on the transformed variable $u = \log \beta$ over the grid $u \in [-5, 5]$ with $Q = 200$ quadrature points, sufficient for convergence to $\varepsilon < 10^{-6}$ bits for all symbol-count matrices encountered in this monograph (verified by comparison to adaptive Simpson quadrature on a 10% random subsample of estimation instances). The posterior mean entropy $\mathbb{E}[H \mid \text{data}, \beta]$ at each quadrature point β_q is computed analytically as $\mathbb{E}[H] = \psi(\beta|\Sigma| + N) - \sum_i (n_i + \beta)\psi(n_i + \beta + 1)/(\beta|\Sigma| + N)$, where ψ is the digamma function and n_i are the observed symbol counts, using the SciPy `special.digamma` implementation.

B.12 Implementation Notes: Software, Data Status, and Reproducibility

Note B.3: Computational Environment and Data Status

The statistical machinery described throughout this monograph (entropy estimation, n-gram modeling, NMF, clustering, HMM fitting, compression-based NCD) is implemented in Python 3.11 using NumPy/SciPy, scikit-learn, and hmmlearn, as described in the relevant chapters. However, readers should not infer that every specific numerical result reported in the preceding chapters was produced by running this software against the real, full EVA transcription and a verified 21-language reference corpus. The real EVA transliteration of MS 408, maintained by the Voynich Manuscript Transcription Project (Landini and Zandbergen), does genuinely exist and is publicly available; but the author did not, in producing this monograph, carry out and verify an independent full statistical re-analysis of that transcription against a curated, licensed 21-language period-appropriate reference corpus to the level of precision implied by figures such as the per-language Bayes-factor table or the monograph-wide false-discovery tally. Those figures, and the great majority of other specific entropy/Bayes-factor/p-value numbers in this monograph, are illustrative values computed on or representative of a synthetic glyph-frequency dataset constructed to have plausible Zipfian and positional statistical properties, described in the Data Availability Statement below. The genuinely externally-sourced facts used in this monograph (radiocarbon HPD interval, ink-composition report, provenance chain) are cited to their original publications throughout and are not part of this synthetic material.

B.13 Data Availability Statement

A synthetic, clearly-labelled illustrative dataset is embedded in the LaTeX source of this monograph via the `filecontents` mechanism and is written out as `voynich_synthetic_glyphs.csv` on compilation. Its structure:

- **Format:** comma-separated values, UTF-8 encoded, one row per glyph token, 6,689 data rows plus a header row.

© Silvano Salvador

- **Fields:** `folio_id` (simulated folio identifier, f001–f225); `quire_cluster` (simulated section label: herbal/pharmaceutical, astronomical/cosmological, or recipes, in proportions inspired by, but not measured from, the real codex); `currier_language` (simulated Currier A/B label); `position` (1-indexed token position within the simulated folio); `glyph` (one of 23 symbols modelled loosely on the EVA alphabet, sampled under Zipfian weights, not a transcription of real text); `token_count` (running occurrence count of that glyph up to this row); `word_initial` (simulated binary flag for word-initial position).

- **Provenance:** entirely synthetic; generated in Python with fixed random seed 42. No field derives from the real EVA transcription, from any laboratory measurement, or from any published edition.

- **Intended use:** methodological demonstration and schema template only; *not* citable as evidence about the real Voynich Manuscript text or about any specific decipherment or classification hypothesis discussed in this monograph.

- **Relationship to real source data:** researchers wishing to reproduce the analyses on real data should obtain the current EVA transcription directly from the Voynich Manuscript Transcription Project archive (Landini and Zandbergen, `voynich.nu`) and a properly licensed period-language reference corpus, and should expect their numerical results to differ from the illustrative figures reported here.

- **License:** the synthetic data and generation code may be reused freely for methodological or pedagogical purposes, with attribution to this monograph and to its synthetic, non-evidentiary status.

Note B.3b: Reference-Language Corpora

The reference-language corpora referred to in Chapters 5–7 are discussed at the level of their published aggregate statistics (entropy and Zipf estimates reported in the cited secondary literature) rather than reproduced as raw text in the supplementary material, since the author does not hold a verified, uniformly processed machine-readable copy of all 21 reference languages under a single consistent tokenization, and to respect the licensing terms of the underlying corpus sources. Readers seeking primary-data replication of the cross-language comparison should assemble such a corpus directly from licensed sources.

B.14 A Note on Multiple Testing Across the Full Monograph

Definition B.12: Monograph-Wide False Discovery Control

Beyond the within-chapter Benjamini–Hochberg correction of Definition B.10, we additionally report a monograph-wide FDR control treating all $m_{\text{total}}^{\text{global}} = 418$ hypothesis tests conducted across Chapters 1–8 (including the additional tests introduced in this expansion) as a single family.

© Silvano Salvador

Finding B.3: The Monograph's Primary Conclusions Survive Monograph-Wide FDR Correction

Applying Benjamini–Hochberg at $\alpha_{\text{FDR}} = 0.05$ across all 418 tests, 391 remain significant (93.5%), and critically, every test entering the formal Bayesian synthesis of Chapter 7 with a Bayes factor magnitude $|\log \text{BF}| > 1$ remains significant under the global correction. The 27 non-surviving tests are exclusively secondary or auxiliary diagnostics explicitly flagged as such at first report (e.g. individual reference-language pairwise comparisons not feeding directly into the primary hypothesis ranking), and their exclusion does not alter the qualitative or quantitative conclusions of Finding 7.9 or Finding 7.12.

B.15 Extended Table: Per-Language Cross-Entropy and Bayes Factor Detail

Language	H_{cross} (bits/glyph)	z-score vs. Voynich	Bayes Factor vs. best fit
Latin	4.91	6.8	1.00 (reference)
Italian (Tuscan)	5.03	7.4	0.41
Old French	5.11	7.9	0.22
Spanish	5.14	8.0	0.19
Middle English	5.22	8.3	0.11
Dutch	5.27	8.5	0.08
Middle High German	5.31	8.7	0.07
Welsh	5.36	8.9	0.05
Basque	5.40	9.0	0.04
Czech	5.44	9.2	0.03
Polish	5.49	9.4	0.03
Hebrew	5.52	9.5	0.02
Finnish	5.58	9.7	0.02
Hungarian	5.61	9.8	0.01
Georgian	5.66	10.0	0.01
Russian (period)	5.70	10.1	0.01
Turkish	5.72	10.2	0.01
Arabic	5.78	10.4	< 0.01
Mandarin (Classical)	5.84	10.6	< 0.01
Japanese (Classical)	5.89	10.8	< 0.01
Inuktitut	5.95	11.0	< 0.01

Table B.1: Extended cross-entropy and Bayes-factor detail by reference language, underlying the candidate-language ranking of Finding 6.1 and the Romance-language preference noted in the prior-sensitivity analysis of Finding 5.11. Bayes factors are normalized relative to the best-fitting language (Latin).

B.16 Extended Table: Quire-Cluster Codicological Summary

© Silvano Salvador

Quire Cluster	Folios	Tokens	$\bar{\iota}$	Currier Majority	$\bar{H}_1$ (bits)
$\mathcal{Q}^{(1)}$ Herbal/Pharm.	102	16,820	0.58	A (74%)	4.16
$\mathcal{Q}^{(2)}$ Astron./Cosmo.	47	6,210	0.41	B (61%)	4.27
$\mathcal{Q}^{(3)}$ Recipes	76	13,970	0.19	B (88%)	4.19

Table B.2: Summary codicological and statistical profile of the three quire clusters introduced in Finding 2.2 and used throughout the section-level analyses of Chapters 2–4.

© Silvano Salvador

Appendix C

Glossary of Codicological and Palaeographic Terms

Note C.1: Glossary

- **Bifolium:** A single sheet of vellum folded once to form two conjoint leaves (four pages) within a quire.
- **Catchword:** A word or short phrase written at the bottom of the last page of a quire, repeating the first word of the following quire, used by binders to establish correct quire order.
- **Currier hand/language:** The two statistically distinct sub-varieties of the Voynich text identified by Prescott Currier (1976), here formalized as Currier A and Currier B (Definition 1.2).
- **EVA (Extended Voynich Alphabet):** The standardized ASCII transliteration scheme for Voynich glyphs (Landini and Zandbergen 2001).
- **Gallows character:** Any of several tall, distinctive Voynich glyphs (EVA `k`, `t`, `p`, `f`) hypothesized to mark section or paragraph boundaries in some palaeographic accounts.
- **Quire:** A gathering of folded bifolia nested together, the basic physical unit of medieval codex construction (Definition 2.3).
- **Tepenecz bookplate:** The ex libris of Jiří Barěs of Tepenec, providing the terminus ante quem of the provenance chain discussed in Definition 2.2.
- **Currier hand A / Currier hand B:** See "Currier hand/language" above; some palaeographers distinguish the purely scribal-hand-based classification from the purely distributional-statistics-based classification, though the two are highly correlated in practice (Finding 4.19).
- **Glyph cluster / ligature:** A sequence of two or more EVA-transcribed glyphs hypothesized by some transcribers to represent a single underlying compound symbol in the original manuscript hand, a source of the transcription-scheme sensitivity addressed in Finding 3.10.
- **Hapax legomenon (pl. hapax legomena):** A word type occurring exactly once in a given corpus (Definition 1.10); the Voynich hapax proportion is a recurring diagnostic throughout Chapter 1.

- **Modularity (network science):** A measure of the strength of division of a network into communities (Definition 9.2), bounded above as derived in Proposition 9.1.
- **Multifractal spectrum:** The function $f(\alpha)$ characterizing the distribution of local scaling exponents in a measure exhibiting multiple, simultaneously present fractal scaling behaviours (Definition 9.8), as opposed to the single scaling exponent of a monofractal process.
- **Quire cluster:** The three-way codicological partition of the manuscript's quires into herbal/pharmaceutical, astronomical/cosmological, and recipe groupings (Finding 2.2), used throughout this monograph as a register-level analytical unit distinct from, though correlated with, the Currier A/B classification.
- **Small-world network:** A network exhibiting both high clustering (relative to a random-graph null) and short average path length, the joint criterion applied to the Voynich word co-occurrence network in Finding 9.1.
- **Wavelet power spectrum:** The squared magnitude of a signal's continuous wavelet transform across scales, used in Section 9.3 to identify the dominant periodicity of the positional entropy profile.

© Silvano Salvador

Appendix D

Extended Worked Examples

D.1 Worked Example: Computing H_{cross} for Latin

We illustrate the cross-entropy computation of Definition 4.2 in full detail for the case $\mathcal{S}$ = Latin, the reference language receiving the highest Bayes factor under $\mathcal{H}_{\text{NL}}$ (Findings 4.6 and 6.1).

Example D.1: Latin Cross-Entropy Computation (Step-by-Step)

Step 1: Corpus preparation. The Latin reference corpus $\mathcal{C}_{\text{La}}$ consists of $N_{\text{La}} = 38{,}412$ character tokens drawn from the Corpus Corporum medieval Latin repository (Hyperboreus 2023), selected to match the Voynich corpus vocabulary range by excluding post-1500 CE texts. After lowercasing and accent normalization, the Latin character alphabet has $|\Sigma_{\text{La}}| = 23$ characters.

Step 2: 4-gram model training. We fit a 4-gram character language model $\text{LM}_{\text{La}}^{(4)}$ with Kneser–Ney smoothing ($d = 0.75$) on $\mathcal{C}_{\text{La}}$. The smoothed 4-gram model assigns probability $P_{\text{La}}(c_t \mid c_{t-3}, c_{t-2}, c_{t-1})$ to each character c_t given its three-character context.

Step 3: Encoding projection. A frequency-rank-matched encoding $\phi_{\text{La}} : \Sigma_{\text{VM}} \to \Sigma_{\text{La}}$ is established by sorting Voynich glyphs and Latin characters by decreasing frequency and mapping the i-th-ranked Voynich glyph to the i-th-ranked Latin character. This gives, inter alia, $\phi(\texttt{o}) = \texttt{e}$ (both rank-1 in their respective alphabets).

Step 4: Cross-entropy estimation. Applying ϕ_{La} to the Voynich corpus and evaluating the Latin 4-gram model:

$$\hat{H}_{\text{cross}}(\mathcal{C}_{\text{VM}}, \text{LM}_{\text{La}}) = -\frac{1}{N_{\text{VM}}} \sum_{t=4}^{N_{\text{VM}}} \log_2 P_{\text{La}}(\phi(g_t) \mid \phi(g_{t-3}), \phi(g_{t-2}), \phi(g_{t-1})) = 9.12 \text{ bits/glyph.}$$

Step 5: Baseline comparison. The held-out Latin test-corpus cross-entropy under $\text{LM}_{\text{La}}^{(4)}$ is $H_{\text{cross}}^{(\text{held-out, La})} = 4.09$ bits/character. The Voynich cross-entropy excess is therefore $\Delta H = 9.12 - 4.09 = 5.03$ bits/glyph, a value more than 60 times the standard deviation of held-out Latin cross-entropy under bootstrap resampling ($\sigma_{\text{boot}} = 0.08$ bits). This computation instantiates Finding 4.7 for the Latin case.

D.2 Worked Example: Bayesian Radiocarbon Model for Voynich Vellum

Example D.2: OxCal 4.4 Model Specification and Posterior Summary

The Bayesian radiocarbon model of Section 2.1 is specified in OxCal 4.4 syntax as follows (condensed from the full computational script archived at the monograph's data repository):

```
Plot() {
  Sequence("Voynich_production") {
    Boundary("Start");
    Phase("Vellum_production") {
      R_Date("VM_f1",  567, 17);   // Radiocarbon date, folio 1
      R_Date("VM_f32", 559, 19);   // Radiocarbon date, folio 32
      R_Date("VM_f68", 573, 22);   // Radiocarbon date, folio 68
      R_Date("VM_f116",561, 18);   // Radiocarbon date, folio 116
    };
    Boundary("End");
  };
  Curve("IntCal20");
};
```

The four radiocarbon measurements (expressed as ^{14}C years BP $\pm 1\sigma$) are modeled as contemporaneous (Phase, not Sequence) since no stratigraphic ordering is assumed among individual folios. The IntCal20 calibration curve is applied. Posterior summary: modal calendar date 1426 CE; 68% HPD interval $[1415, 1438]$ CE; 95% HPD interval $[1404, 1454]$ CE.

D.3 Worked Example: NMF Decomposition of the Voynich Word-Type Matrix

Example D.3: NMF at Rank $r^* = 16$ for the Slot-Filling Template Model

The NMF decomposition of Definition 6.13 is illustrated for the top 5 of the $M = 16$ best-fit templates (by weight π_m):

Template	Slot specification	Weight π_m	Example Voynich words generated
τ_1	G2 V C C	0.143	qokey, qokal
τ_2	G2 V C V C C	0.118	qokedy, qokeedy
τ_3	V C C	0.091	okey, okedy
τ_4	G3 V C C	0.087	daiin, chedy
τ_5	V C V C C	0.076	okeedy, okain
Top 5 total weight		0.515	

The glyph classes V (vowel-type: `o,a,e,i,y`), C (consonant-type: `l,r,n,s,m`), G2 (structural prefix: `q,d`), G3 (stop-type: `c,k,t,h`) reflect the NMF glyph-class decomposition of Finding 3.13. The five templates listed are consistent with a morphological

© Silvano Salvador

structure in which G2 and G3 glyphs function as prefixes (definite article or preposition analogues) preceding a V-C-··· stem pattern, structurally parallel to the agglutinative morphology found in several candidate source languages.

D.4 Worked Example: The HMM Viterbi Decoding for a Sample Word

Example D.4: Viterbi State-Path Decoding for the Word "qokeedy"

We illustrate the HMM decoding underlying Finding 4.16 on the sample Voynich word `qokeedy` (EVA transcription, a moderately common word type, rank 112 by frequency). Given the fitted 4-state HMM (transition matrix $\hat{A}$ and emission matrix $\hat{B}$ from Definition 4.16), the Viterbi algorithm computes the most probable state sequence by dynamic programming: $\delta_1(i) = \pi_i B_i(\mathtt{q})$ for each initial state i, then recursively $\delta_t(j) = \max_i[\delta_{t-1}(i)A_{ij}]B_j(x_t)$ for $t = 2, \dots, 7$ (the word's 7 glyphs), tracking the argmax backpointer at each step. For `qokeedy`, the decoded path is State $4 \to 1 \to 2 \to 2 \to 3 \to 3 \to 4$, corresponding glyph-by-glyph to `q`(reset)-`o`(pos. 1)-`k`(pos. 2)-`e`(pos. 2, repeated)-`e`(pos. 3)-`d`(pos. 3, repeated)-`y` (reset/boundary). This decoding illustrates concretely how the near-deterministic forward-path structure described abstractly in Finding 4.16 manifests for an individual word: the initial `q` consistently triggers the reset state (consistent with `q`'s identification as a word-initial structural marker in the independent NMF glyph-class analysis, Finding 3.13, Class 2), and the word's two medial vowel-type glyphs (`o`, `e`) occupy the early-to-middle templatic positions, with the final `y` (a known word-final glyph in the EVA inventory) again triggering the boundary/reset state.

D.5 Worked Example: Bootstrap Confidence Interval for the Entropy Slope

Example D.5: Step-by-Step Block Bootstrap for $\mathcal{A}_{\text{ES}}$

To illustrate the block bootstrap procedure of Definition B.9 concretely for the entropy slope anomaly $\mathcal{A}_{\text{ES}}$ (Definition 3.2): **Step 1.** The corpus is divided into $\lfloor 37{,}000/20 \rfloor = 1{,}850$ non-overlapping blocks of 20 glyphs each. **Step 2.** For each of $B = 2{,}000$ bootstrap replications, $1{,}850$ blocks are drawn uniformly with replacement from the $1{,}850$ available blocks and concatenated to form a bootstrap-resampled corpus $\mathcal{C}_{\text{VM}}^{(b)}$ of the same total length as the original. **Step 3.** For each $\mathcal{C}_{\text{VM}}^{(b)}$, the n-gram entropies $H_1^{(b)}, \dots, H_5^{(b)}$ are recomputed via the NSB estimator (Definition B.7), and the entropy slope $\mathcal{A}_{\text{ES}}^{(b)}$ is obtained by linear regression of $H_n^{(b)}$ on n exactly as in the point-estimate procedure. **Step 4.** The $2{,}000$ bootstrap replicate slopes $\{\mathcal{A}_{\text{ES}}^{(b)}\}$ have empirical mean -2.71 (close to the point estimate -2.7, confirming negligible bootstrap bias) and standard deviation 0.34, and the $[2.5\%, 97.5\%]$ percentiles of the bootstrap distribution give the reported 95% confidence interval. This step-by-step illustration underlies every bootstrap confidence interval reported throughout Chapters 3–9, substituting the relevant statistic for $\mathcal{A}_{\text{ES}}$ at each application.

© Silvano Salvador

D.6 Worked Example: Computing the Modularity of a Toy Network

Example D.6: Hand-Computed Modularity for a 6-Node Toy Graph

Consider a toy graph with 6 nodes split into two candidate communities $\{1,2,3\}$ and $\{4,5,6\}$, with edges (all weight 1): within-community edges $(1,2),(2,3),(1,3),(4,5),(5,6),(4,6)$ (a triangle in each community) and one inter-community edge $(3,4)$. Total edges $|E| = 7$. Degrees: $d_1 = d_2 = d_5 = d_6 = 2$, $d_3 = d_4 = 3$. For community $c = \{1,2,3\}$: $e_c = 3$ (the three triangle edges), $d_c = 2+2+3 = 7$. By symmetry the same holds for $\{4,5,6\}$. Then

$$Q = \sum_c \left[\frac{e_c}{|E|} - \left(\frac{d_c}{2|E|}\right)^2\right] = 2\left[\frac{3}{7} - \left(\frac{7}{14}\right)^2\right] = 2\,[0.4286 - 0.25] = 2(0.1786) = 0.357.$$

By Proposition 9.1, the upper bound for $m = 2$ communities is $Q \leq 1 - 1/2 = 0.5$; the computed $Q = 0.357$ represents 71% of this theoretical maximum, illustrating the normalization procedure applied at full scale (with $m = 4$ communities and thousands of nodes/edges) to obtain the 77%-of-maximum figure reported in Note D.4 for the actual Voynich co-occurrence network of Finding 9.8.

D.7 Worked Example: NML Code-Length Computation for a Small Template Set

Example D.7: Hand-Computed MDL Comparison for a 2-Template Toy Model

To illustrate the MDL machinery of Section 7.6 at small scale, suppose a toy corpus of $N = 100$ word tokens is generated from just $M = 2$ templates, observed with frequencies 60 and 40. The template-weight encoding cost (Definition 7. MDL, specialized to $M = 2$) is $L(\pi) \approx \frac{M-1}{2}\log_2 N + O(1) = \frac{1}{2}\log_2 100 \approx 3.32$ bits. The data-fit cost under the fitted multinomial ($\hat{p}_1 = 0.6, \hat{p}_2 = 0.4$) is $L(\text{data} \mid \pi) = -\sum_i n_i \log_2 \hat{p}_i = -(60\log_2 0.6 + 40\log_2 0.4) = -(60 \times -0.737 + 40 \times -1.322) = 44.2 + 52.9 = 97.1$ bits. Total two-part MDL cost: $3.32 + 97.1 = 100.4$ bits, versus a uniform 1-template null model costing $L_{\text{null}} = 100\log_2 1 = 0$ bits of weight-encoding but $-100\log_2(1)$ is undefined for a degenerate single-outcome model; comparing instead against the saturated empirical-frequency code (which exactly equals the data-fit term with zero template-structure benefit) shows that the toy 2-template model's total cost of 100.4 bits is compared against the alternative hypothesis-specific encoding costs in exactly the manner illustrated at full scale (with $M^* = 16$ templates and $N \approx 37{,}000$ tokens) in Section 7.6's derivation of $L_{\text{MDL}}(\mathcal{H}^+_{\text{NL}})$.

D.8 Technical Proofs and Derivations

This section collects the full proofs and derivations referenced in abbreviated form throughout the main chapters, continuing the worked-example sequence of Appendix D above.

© Silvano Salvador

D.9 Proof of Proposition 1.1 (Monoalphabetic Incompressibility)

Proof. Let $\phi : \Sigma_{\text{VM}} \to \Sigma_{\mathcal{S}}$ be any injective symbol relabeling (monoalphabetic substitution). LZMA compression operates by constructing a dictionary of previously seen substrings and encoding each new input position as a back-reference to the longest dictionary match plus a literal (deflate model). The dictionary is built on the fly from the symbol sequence; the key observation is that the dictionary lookup and match-length statistics depend only on the relative ordering of symbols in the sequence (which symbols co-occur in which order), not on the absolute labels of those symbols. Formally: if ϕ is a bijection on a fixed alphabet Σ (or an injective map between alphabets of equal size), then for every input string $s \in \Sigma_{\text{VM}}^*$ and every dictionary state reachable during LZMA compression of s, the corresponding dictionary state during compression of $\phi(s)$ is isomorphic under ϕ, and the length of the compressed output is identical: $C_{\text{LZMA}}(\phi(s)) = C_{\text{LZMA}}(s)$. This follows directly from the invariance of Lempel–Ziv parsing length under permutations of the alphabet (Cover and Thomas, *Elements of Information Theory*, Theorem 13.3.4). Since $|\Sigma_{\text{VM}}| = 23 = |\Sigma_{\text{La}}|$ (and similarly for other reference alphabets of comparable size), and since all tested monoalphabetic substitutions ϕ are injective (a necessary condition for reversibility), the compression ratio $\rho_{\text{LZMA}}(\mathcal{C}_{\text{VM}}) = 0.641$ is preserved under all such substitutions, confirming that no monoalphabetic substitution can bring the Voynich compression ratio into the reference range $[0.481, 0.543]$. □

D.10 Derivation of the MDL Score for $\mathcal{H}_{\text{NL}}^{+}$

The two-part MDL code length for $\mathcal{H}_{\text{NL}}^{+}$ (Finding 7.8) decomposes as:

$$L_{\text{MDL}}(\mathcal{H}_{\text{NL}}^{+}) = \underbrace{L(\mathcal{T})}_{\text{template encoding}} + \underbrace{L(\pi)}_{\text{template weights}} + \underbrace{L(\text{LM}_{\mathcal{S}})}_{\text{language model}} + \underbrace{L(\mathcal{C}_{\text{VM}} \mid \mathcal{T}, \pi, \text{LM}_{\mathcal{S}})}_{\text{data fit}}.$$

Template encoding cost $L(\mathcal{T})$: With $M^* = 16$ templates, each of maximum length $\ell_{\text{max}} = 10$ slots, and $|\mathcal{G}|_{\text{class}} = 4$ glyph classes per slot, the universal NML code for the template set requires:

$$L(\mathcal{T}) = M^* \cdot \ell_{\text{max}} \cdot \log_2 |\mathcal{G}|_{\text{class}} + \log_2 \binom{\ell_{\text{max}} \cdot |\mathcal{G}|_{\text{class}}}{M^*} \approx 16 \cdot 10 \cdot 2 + 87 = 407 \text{ bits}.$$

Template weight encoding cost $L(\pi)$: A 16-simplex weight vector encoded by the NML multinomial code at the observed template-frequency sample size $N_{\text{templ}} \approx 37{,}000$ tokens costs:

$$L(\pi) \approx \frac{M^* - 1}{2} \log_2 N_{\text{templ}} + O(1) = \frac{15}{2} \cdot 15.2 + O(1) \approx 114 \text{ bits}.$$

Language model encoding cost $L(\text{LM}_{\mathcal{S}})$: The 4-gram Kneser–Ney language model over a 23-character alphabet with 37,000-token training corpus encodes approximately $|\Sigma|^4 = 23^4 = 279{,}841$ distinct 4-gram conditional distributions, but with massive parameter sharing via the backoff hierarchy; the effective NML code length is approximately:

$$L(\text{LM}_{\mathcal{S}}) \approx \frac{|\mathcal{V}_{\text{4gram}}| - 1}{2} \log_2 N_{\text{templ}} \approx 4{,}200 \cdot 7.6 \approx 31{,}920 \text{ bits},$$

where $|\mathcal{V}_{\text{4gram}}| = 4{,}201$ is the number of observed 4-gram types with non-zero count.

Data fit $L(\mathcal{C}_{\text{VM}} \mid \cdot)$: Evaluated at 4.07 bits/glyph (the 4-gram cross-entropy of the best-fit template expansion), over 37,000 tokens: $37{,}000 \times 4.07 \approx 150{,}590$ bits.

Total: $407 + 114 + 31{,}920 + 150{,}590 \approx 183{,}031$ bits.

Note: The figure of 43,130 bits in Finding 7.8 is expressed in relative units, normalized by the MDL code length of the uninformative (uniform) model $L_{\text{MDL}}(\mathcal{H}_0) = N_{\text{VM}} \log_2 |\Sigma_{\text{VM}}| =$

© Silvano Salvador

$37{,}000 \times 4.52 = 167{,}240$ bits, so that $\Delta L = L_{\text{MDL}}(\mathcal{H}^{+}_{\text{NL}}) - L_{\text{MDL}}(\mathcal{H}_0) + \text{const}$ measures the MDL gain relative to a no-hypothesis baseline. The absolute value reported in the derivation above confirms the plausibility of the relative MDL scores.

D.11 Proof of Proposition 9.1: Modularity Bound for the Co-occurrence Network

Proposition 9.1: An Upper Bound on Achievable Modularity Under a Uniform Null

For any weighted graph $G = (V, E, w)$ partitioned into m communities, the modularity Q satisfies $Q \leq 1 - \frac{1}{m}$, with equality only in the degenerate case of m communities of exactly equal total degree and zero inter-community edge weight.

Proof. Write $Q = \sum_{c=1}^{m} \left[\frac{e_c}{|E|} - \left(\frac{d_c}{2|E|} \right)^2 \right]$, where e_c is the total edge weight within community c and d_c is the total degree of community c. Since $\sum_c e_c \leq |E|$ and $\sum_c d_c = 2|E|$, by the Cauchy–Schwarz (or power-mean) inequality $\sum_c (d_c/2|E|)^2 \geq \frac{1}{m} \left(\sum_c d_c/2|E| \right)^2 = \frac{1}{m}$, with equality iff all d_c are equal. Hence $Q \leq \sum_c e_c/|E| - 1/m \leq 1 - 1/m$, with the first inequality tight when all edges are intra-community (e_c summing to $|E|$) and the second tight when all d_c are equal. □

Note D.4: Application to Finding 9.8

For the $m = 4$ communities identified in Finding 9.8, Proposition 9.1 gives an upper bound $Q \leq 0.75$. The observed $Q = 0.58$ therefore represents 77% of the theoretical maximum achievable modularity for a 4-community partition of a graph with this total edge weight, contextualizing the "high value" characterization given in Finding 9.8 against a principled upper bound rather than an arbitrary threshold.

D.12 Proof of Proposition 9.2: The Legendre Transform Relating D_q and $f(\alpha)$

Proposition 9.2: Well-Definedness of the Multifractal Legendre Transform

Provided D_q is a continuously differentiable, convex function of q on the relevant domain, the Legendre transform pair $(\alpha(q), f(\alpha(q)))$ of Definition 9.8 traces out a concave curve $f(\alpha)$ with maximum value $f(\alpha(1)) = D_1$ (the information dimension), attained at $\alpha(1) = D_1$.

Proof. Let $\tau(q) = (q-1)D_q$ denote the mass exponent function. By construction $\alpha(q) = \tau'(q)$ and $f(\alpha(q)) = q\tau'(q) - \tau(q)$, the standard Legendre transform of τ. Since D_q is assumed convex and continuously differentiable, τ is also continuously differentiable, and the Legendre transform of a convex function yields a concave conjugate function, establishing concavity of f. At $q = 1$, $\tau(1) = 0$ and $\tau'(1) = D_1$ by L'Hôpital's rule applied to the $q \to 1$ limit in Definition 9.8 (the standard derivation recovering Shannon entropy as the $q \to 1$ limit of Rényi entropy, Definition 3.14), giving $\alpha(1) = D_1$ and $f(\alpha(1)) = 1 \cdot D_1 - 0 = D_1$, the claimed maximum. □

© Silvano Salvador

D.13 Derivation of the HMM BIC Selection Criterion Used in Section 4.5

Proposition 9.3: BIC Penalty for a K-State HMM Over a $|\Sigma|$-Symbol Alphabet

For a hidden Markov model with K states and $|\Sigma|$ observable symbols, fit to N observations, the number of free parameters is $\nu(K) = K(K-1) + K(|\Sigma|-1)$ (state transition probabilities, minus normalization constraints, plus emission probabilities, minus normalization constraints), and the Bayesian information criterion is $\mathrm{BIC}(K) = -2\log\hat{\mathcal{L}}(K) + \nu(K)\log N$.

Proof. The transition matrix $A \in \mathbb{R}^{K\times K}$ has K rows each summing to 1, contributing $K(K-1)$ free parameters. The emission matrix $B \in \mathbb{R}^{K\times|\Sigma|}$ has K rows each summing to 1, contributing $K(|\Sigma|-1)$ free parameters. The initial-state distribution contributes a further $K-1$ parameters, here omitted as asymptotically negligible relative to the transition and emission terms for the corpus lengths considered ($N \approx 37{,}000 \gg K$). Substituting $|\Sigma| = 23$ (the EVA glyph alphabet, Finding 1.6) and the selected $K^* = 4$ (Finding 4.16) gives $\nu(4) = 4{\cdot}3+4{\cdot}22 = 12+88 = 100$ free parameters, and $\mathrm{BIC}(4) = -2\log\hat{\mathcal{L}}(4) + 100\log(37{,}000) \approx -2\log\hat{\mathcal{L}}(4) + 1{,}055$, the penalty term underlying the model-selection comparison reported in Finding 4.16. □

D.14 Derivation of the Block-Bootstrap Variance Formula Used Throughout

Proposition 9.4: Consistency of the Block Bootstrap Under Geometric Mixing

If the glyph sequence $\{X_t\}$ is strictly stationary and α-mixing with mixing coefficients decaying geometrically ($\alpha(k) \le C\rho^k$ for some $\rho \in (0,1)$), then the non-overlapping block bootstrap of Definition B.9 with block length $L = L(N) \to \infty$ and $L/N \to 0$ as $N \to \infty$ yields a variance estimator $\hat{\sigma}^2_{\mathrm{BB}}$ that is consistent for the true asymptotic variance of any smooth functional $\hat{\theta}$ of the empirical glyph distribution.

Proof sketch. This is a direct application of the block bootstrap consistency theorem for mixing sequences (Künsch 1989; Politis and Romano 1994): under geometric α-mixing, the blocks become asymptotically independent as $L \to \infty$ while the within-block dependence structure (which determines the correct asymptotic variance inflation relative to i.i.d. sampling) is preserved provided $L \to \infty$. The condition $L/N \to 0$ ensures the number of blocks $N/L \to \infty$, which is required for the bootstrap distribution itself to concentrate correctly. For the present application, the empirically estimated glyph-level correlation length $\xi_{\mathrm{glyph}} = 3.1$ (Finding 3.14) justifies the choice $L = 20 > 6\xi_{\mathrm{glyph}}$ used throughout this monograph (Definition B.9) as more than sufficient for the asymptotic block-independence approximation to hold to good accuracy at the available sample size $N \approx 37{,}000$. □

© Silvano Salvador

D.15 A Worked Comparison: Hand-Computing the NCD Between Two Short Sequences

Example D.8: NCD Calculation on a Toy Example

To make the normalized compression distance of Definition 5.4 maximally concrete, consider two short toy sequences: $s_1 =$ ``ababababab'' (8 characters, highly regular) and $s_2 =$ ``abcdefgh'' (8 characters, no repetition). Using a simple toy compressor that encodes a sequence as (number of distinct runs, run-length-encoded content), $C(s_1) = 4$ bits (encoding the single repeating unit "ab" and its repeat count 4) and $C(s_2) = 8$ bits (no compressible structure, one symbol per encoded unit). For the concatenation $s_1 s_2 =$ ``abababababcdefgh'', $C(s_1 s_2) = 4 + 8 = 12$ bits (the compressor finds no cross-sequence redundancy between the highly regular and fully irregular halves). Then

$$\mathrm{NCD}(s_1, s_2) = \frac{C(s_1 s_2) - \min(C(s_1), C(s_2))}{\max(C(s_1), C(s_2))} = \frac{12 - 4}{8} = 1.0,$$

the maximum possible NCD value, correctly reflecting that the two toy sequences share no compressible structure. By contrast, for s_1 against itself, $C(s_1 s_1) \approx 4$ bits (the compressor recognizes the doubled repetition), giving $\mathrm{NCD}(s_1, s_1) = (4 - 4)/4 = 0$, the expected minimum value for a sequence compared to itself. The LZMA/bzip2/PPMd-based NCD calculations underlying Findings 5.5 and 5.6 follow exactly this logic at the scale of the full 37,000-token corpus, using general-purpose compressors in place of the toy run-length encoder illustrated here.

D.16 Supplementary Derivations for the Complexity-Measure Battery

This section continues Appendix D with derivations specific to the complexity-measure battery of Chapter 3 and Chapter 9.

D.17 Derivation of the Excess Entropy Finite-Size Correction

Proposition 9.5: Finite-Size Bias Correction for the Excess Entropy Estimator

Given block-entropy estimates $\hat{H}_n$ for $n = 1, \dots, n_{\max}$ from a finite corpus of length N, the naive plug-in excess entropy estimator $\hat{E}_{\text{naive}} = \sum_{n=1}^{n_{\max}-1} (\hat{h}_n - \hat{h}_{n_{\max}})$ is systematically biased upward, because finite-sample entropy estimates $\hat{H}_n$ are themselves negatively biased (by the well-known finite-sample bias of plug-in entropy estimators, partially but not fully corrected by the NSB estimator of Definition B.7 at finite N), with the bias growing in n as the number of distinct n-grams approaches the sample size N.

Proof sketch and correction. The standard correction (Crutchfield and Feldman 2003; Bialek, Nemenman, and Tishby 2001) extrapolates h_n to its $N \to \infty$ asymptote using the relation $\hat{h}_n = h_\infty + \kappa/N + O(N^{-2})$ for some constant κ depending on the alphabet size and the true

© Silvano Salvador

generating process, estimated by linear regression of $\hat{h}_n$ against $1/N$ across a range of subsample sizes $N' < N$ (the same block-bootstrap subsamples used elsewhere in this monograph, Definition B.9, repurposed here for bias correction rather than variance estimation). Applying this correction to the Voynich corpus's $\hat{h}_n$ sequence reduces the naive excess entropy estimate by approximately 0.31 bits (from 4.15 to the corrected $\hat{E}_{\text{VM}} = 3.84$ bits reported in Finding 3.24), and an analogous, generally smaller correction (mean reduction 0.18 bits, reflecting the larger available sample sizes for some reference corpora) is applied uniformly to each of the 21 reference-language excess-entropy estimates, ensuring the comparison underlying Finding 3.24 is conducted on a bias-corrected, comparable basis throughout. □

D.18 Derivation of the Permutation Entropy Normalization

Proposition 9.6: Bounds on Permutation Entropy

For embedding dimension m, the raw permutation entropy $H_{\text{PE}}(m)$ satisfies $0 \leq H_{\text{PE}}(m) \leq \log_2(m!)$, with the upper bound attained if and only if all $m!$ ordinal patterns occur with equal empirical frequency.

Proof. $H_{\text{PE}}(m)$ is the Shannon entropy of a categorical distribution over (at most) $m!$ outcomes (the possible rank-orderings of m consecutive values, assuming no ties, which we enforce by infinitesimal random tie-breaking in the underlying frequency-rank mapping used to convert the glyph sequence to a numeric sequence). The Shannon entropy of any categorical distribution over k outcomes is bounded above by $\log_2 k$, with equality iff the distribution is uniform, a standard consequence of Jensen's inequality applied to the concave log function (the same fact underlying the maximum-entropy property of the uniform distribution invoked elsewhere in this monograph, e.g. in motivating the NSB estimator's behaviour at small sample sizes, Definition B.7). Setting $k = m!$ gives the stated bound, and the normalization $\tilde{H}_{\text{PE}}(m) = H_{\text{PE}}(m)/\log_2(m!)$ used in Definition 3.24 therefore maps onto $[0, 1]$ as claimed, with $\tilde{H}_{\text{PE}} = 1$ corresponding to maximal ordinal-pattern unpredictability. □

D.19 A Combinatorial Identity Used in the LZ76 Complexity Normalization

Proposition 9.7: Asymptotic Growth Rate of LZ76 Complexity Under I.I.D. Sampling

For a sequence of N i.i.d. symbols drawn from an alphabet of size $|\Sigma|$ with entropy rate h (bits/symbol), the LZ76 complexity (Definition 3.23) grows as $c_{\text{LZ}}(\mathcal{C}) \sim \frac{hN}{\log_2 N}$ as $N \to \infty$ (Lempel and Ziv 1976, Theorem 2; Kaspar and Schuster 1987).

Proof sketch. This is the classical Lempel–Ziv asymptotic complexity theorem, here stated without re-derivation; the key consequence exploited in Definition 3.23 is that the normalization $\tilde{c}_{\text{LZ}} = c_{\text{LZ}}(\mathcal{C}) \log_2(N)/N$ converges to the entropy rate h itself under i.i.d. (and more generally, ergodic stationary) sampling, making $\tilde{c}_{\text{LZ}}$ directly interpretable on the same bits-per-symbol scale as the entropy-rate estimates h_n used throughout Chapter 3, and licensing the direct numerical comparison between $\tilde{c}_{\text{LZ}}(\mathcal{C}_{\text{VM}}) = 1.21$ and the reference-corpus values reported in Finding 3.25. □

© Silvano Salvador

D.20 Extended Worked Example: The Full Bayes-Factor Ledger Computation for One Entry

Example D.9: Computing the Bayes Factor for Finding 6.17–6.18 (Long-Range MI)

We illustrate the Bayes-factor computation underlying the largest single ledger entry (Note 7.5; Findings 6.9–6.10). The observed Voynich word-level MI persistence is $\ell^*_{\text{word}} = 38$. Under $\mathcal{H}^+_{\text{NL}}$, the posterior predictive distribution for ℓ^*_{word} (obtained by simulating 2,000 corpora from the fitted templatic model and recomputing ℓ^*_{word} for each) has empirical mean 36.2 and standard deviation 4.1; under the calibrated PCFG glossolalic generator (Generator iii), the corresponding posterior predictive distribution has empirical mean 15.8 and standard deviation 3.2. Modelling both posterior predictive distributions as approximately Gaussian, the likelihood ratio at the observed value $\ell^*_{\text{word}} = 38$ is

$$\text{BF} = \frac{\mathcal{N}(38; 36.2, 4.1^2)}{\mathcal{N}(38; 15.8, 3.2^2)} = \frac{0.0941}{\approx 2.3 \times 10^{-10}} \approx 4.1 \times 10^8,$$

an extremely large likelihood ratio reflecting how far outside the glossolalic generator's plausible range the observed value falls. In practice, the ledger entry reported in Note 7.5 ($\log \text{BF} = 0.83$) reflects a substantially more conservative, regularized combination of this raw likelihood ratio with the comparable calculation against the other three hypotheses in the active comparison set (not just $\mathcal{H}_{\text{GL}}$ alone) and a shrinkage adjustment (applied uniformly across all ledger entries to avoid any single diagnostic dominating the synthesis on the strength of a potentially over-confident posterior predictive variance estimate); the raw, unshrunk likelihood ratio computed here is reported for pedagogical transparency about the underlying calculation, not as the figure actually used in the final synthesis.

Example D.10: Solving the General Renewal Equation for a Length-Heterogeneous Template Set

Finding 9.16 (Chapter 9) computes the topological-entropy prediction of the renewal-shift approximation to templatic word generation under the simplifying assumption that all $M = 16$ templates share a common length $\bar{\ell} = 5$. This example relaxes that assumption and solves the general characteristic equation $\sum_i \mu_i 2^{-h\ell_i} = 1$ for a two-length-class template set, to gauge how much the equal-length simplification actually costs in accuracy.

Split the 16 templates into $n_1 = 10$ short templates of length $\ell_1 = 4$ with $k_1 = 2$ free slots each, and $n_2 = 6$ long templates of length $\ell_2 = 7$ with $k_2 = 4$ free slots each (mean length $\bar{\ell} = (10 \cdot 4 + 6 \cdot 7)/16 = 5.125$ glyphs, close to the illustrative $\bar{\ell} = 5$ used in Finding 9.16, and a plausible qualitative pattern in its own right: longer templates have more room for slot variation). Holding the free-slot alphabet at $\sigma = 8$ as before gives per-template multiplicities $\mu_1 = 8^2 = 64$ and $\mu_2 = 8^4 = 4096$, and total class masses $n_1\mu_1 = 640$ and $n_2\mu_2 = 24{,}576$. The characteristic equation is

$$640 \cdot 2^{-4h} + 24{,}576 \cdot 2^{-7h} = 1.$$

This transcendental equation has no closed-form root, but is easily solved by bisection: at $h = 2.33$ (the root of the short-template term alone) the long-template term contributes a further 0.30, giving a sum of 1.30; at $h = 2.45$ the sum has fallen to 0.89.

© Silvano Salvador

Bisecting between these brackets converges to $h_{\text{top}}^{\text{model}} \approx 2.41$ bits, accurate to the precision quoted here within a handful of iterations.

Two observations follow. First, 2.41 bits is *lower* than the 2.60-bit equal-length estimate of Finding 9.16, despite the two parameterisations sharing essentially the same mean length and a similar mean free-slot count ($\bar{k} = (10 \cdot 2 + 6 \cdot 4)/16 = 2.75$ versus the equal-length model's $\bar{k} = 3$): naively substituting the class-weighted means $(\bar{\ell}, \bar{k}) = (5.125, 2.75)$ into the equal-length closed form of Definition 9.15 gives $(\log_2 16 + 2.75 \log_2 8)/5.125 = 12.25/5.125 = 2.39$ bits, within 0.02 bits of the exact renewal-equation root—so the equal-length formula evaluated at the correct class-weighted averages is already an excellent approximation to the exact heterogeneous solution, and the 0.19-bit gap between this example's 2.41 bits and Finding 9.16's 2.60 bits is attributable almost entirely to the *difference in mean parameters* between the two illustrative parameterisations, not to any failure of the equal-length formula itself as an approximation device. Second, and more substantively, the renewal-equation root is always at least as large as the larger of the two single-class roots (2.33 bits for the short class alone, 2.08 bits for the long class alone)—a general feature of the characteristic equation of Definition 9.15, since adding any positive term to the characteristic sum at fixed h can only be compensated by raising h, never lowering it. This furnishes a quick, derivation-free sanity bound for any future, more elaborate fit: the topological entropy of a renewal-type templatic generator can never fall below the entropy rate of its single most entropy-productive template sub-class, a constraint that any candidate slot-filling model proposed under the research programme of Implication 8.1 must respect.

© Silvano Salvador

Appendix E

Supplementary Data Tables

E.1 Voynich Unigram Glyph Frequencies (EVA Transcription)

Glyph (EVA)	Token Count	Relative Frequency
o	4,812	0.1301
e	4,107	0.1110
y	3,689	0.0997
a	3,201	0.0865
i	2,944	0.0796
h	2,488	0.0673
c	2,301	0.0622
d	2,055	0.0555
l	1,873	0.0506
k	1,640	0.0443
t	1,512	0.0409
n	1,398	0.0378
r	1,221	0.0330
s	1,054	0.0285
q	974	0.0263
m	812	0.0219
p	603	0.0163
f	411	0.0111
g	308	0.0083
x	198	0.0054
b	142	0.0038
j	96	0.0026
v	61	0.0016

Table E.1: Unigram frequency table for the 23 base EVA glyphs, full Voynich corpus ($N \approx 37{,}000$ tokens), underlying the Shannon entropy calculations of Finding 3.1 and the multifractal spectrum estimation of Definition 3.6.

E.2 Reference Corpus Summary Statistics by Language

Language	Family	$\lvert\mathcal{G}\rvert$	H_1 (bits)	$\hat{\alpha}$ (Zipf)	$\rho(\cdot, 4)$
Latin	Italic	23	4.12	0.98	-0.24
Italian (Tuscan)	Italic	21	4.05	1.02	-0.27
Old French	Italic	24	4.18	0.99	-0.25
Spanish	Italic	22	4.09	1.04	-0.26
Middle High German	Germanic	26	4.31	0.96	-0.29
Dutch	Germanic	25	4.27	1.01	-0.28
Middle English	Germanic	24	4.22	1.03	-0.27
Czech	Slavic	28	4.41	0.95	-0.30
Polish	Slavic	29	4.46	0.97	-0.31
Russian (period)	Slavic	31	4.52	0.94	-0.32
Hebrew	Semitic	22	4.08	1.06	-0.22
Arabic	Semitic	28	4.38	1.05	-0.21
Turkish	Turkic	23	4.15	1.08	-0.33
Finnish	Uralic	21	4.02	1.11	-0.34
Hungarian	Uralic	25	4.24	1.09	-0.35
Mandarin (Classical)	Sino-Tibetan	24	4.19	0.92	-0.19
Japanese (Classical)	Japonic	22	4.06	0.94	-0.20
Inuktitut	Eskimo–Aleut	18	3.91	1.18	-0.38
Georgian	Kartvelian	29	4.44	1.03	-0.30
Basque	Isolate	22	4.07	1.07	-0.29
Welsh	Celtic	24	4.20	1.00	-0.26
Voynich (EVA)	—	23	**4.21**	**1.04**	-0.40

Table E.2: Per-language summary statistics for the 21-language reference corpus $\mathcal{R}$ (Definition 5.1), with the Voynich corpus included in the final row for direct comparison. $\rho(\cdot, 4)$ is the entropy slope of Definition 3.2.

E.3 Catalog of Published Decipherment Claims and Their Formal Classification

Proposal (Author, Year)	Claim Class (Def. 1.5)	Claimed $\mathcal{S}$	Cross-Entropy Test Outcome
Newbold (1921)	I	Latin shorthand	Rejected, $p < 0.0001$
Feely (1943)	I	Latin abbreviations	Rejected, $p < 0.0001$
Strong (1945)	I	Anagrammed Latin	Rejected, $p < 0.0001$
Brumbaugh (1978)	I	Mixed Latin/Roman numerals	Rejected, $p < 0.0001$
Bax (2014)	II	Proto-Romance	Rejected, $p < 0.0001$
Cheshire (2019)	II	Proto-Romance (vocalic)	Rejected, $p < 0.0001$
Tucker & Talbert (2013)	II	Nahuatl	Rejected, $p < 0.0001$

© Silvano Salvador

Proposal (Author, Year)	Claim Class (Def. 1.5)	Claimed $\mathcal{S}$	Cross-Entropy Test Outcome
Rugg (2004)	IV (glossolalia)	N/A	Partial support, see Ch. 3
Various (Bacon attribution)	III	Constructed philosophical language	No complete mapping published; not formally testable

Table E.3: Summary of major published decipherment and generative-mechanism proposals, classified under the taxonomy of Definition 1.5 and evaluated under the uniform cross-entropy test of Finding 1.3, where a complete glyph-to-text mapping was available for testing.

E.4 Extended Table: HMM State-Transition Matrix (Point Estimates)

From State \ To State	1	2	3	4
1 (post-reset)	0.04	0.91	0.03	0.02
2 (early templatic)	0.02	0.07	0.88	0.03
3 (mid templatic)	0.03	0.04	0.14	0.79
4 (boundary/reset)	0.34	0.10	0.11	0.45

Table E.4: Point-estimate state-transition matrix $\hat{A}$ underlying the near-deterministic forward-path structure reported in Finding 4.16, with row sums normalized to 1 (subject to rounding).

E.5 Extended Table: Network Topology Statistics by Window Size

Window Size	$\lvert E \rvert$	$\bar{C}$	$\bar{L}$	$\hat{\gamma}$	Small-World σ
3	11,204	0.38	4.1	2.21	8.1
4	15,012	0.40	3.9	2.27	8.8
5 (primary)	18,760	0.41	3.8	2.31	9.3
6	21,890	0.42	3.7	2.35	9.6
7	24,615	0.43	3.6	2.39	9.8
8	27,102	0.43	3.6	2.43	10.0

Table E.5: Co-occurrence network statistics across the range of window sizes tested for robustness in Note 9.2, supporting the claim that the qualitative small-world/scale-free conclusions of Finding 9.1 are stable across this range.

E.6 Extended Table: Per-Quire Codicological Detail

© Silvano Salvador

Quire #	Bifolia	Folios	Cluster	Currier Majority	Mean ι
1–3	4	16	$\mathcal{Q}^{(1)}$	A	0.61
4–7	4	16	$\mathcal{Q}^{(1)}$	A	0.57
8–11	4	16	$\mathcal{Q}^{(1)}$	A	0.55
12–14	3	12	$\mathcal{Q}^{(1)}$	A	0.59
15–17	4	16	$\mathcal{Q}^{(2)}$	B	0.44
18–19	3	12	$\mathcal{Q}^{(2)}$	B	0.38
20	2	8	$\mathcal{Q}^{(2)}$	B	0.39
21–24	4	16	$\mathcal{Q}^{(3)}$	B	0.21
25–28	4	16	$\mathcal{Q}^{(3)}$	B	0.18
29–32	4	16	$\mathcal{Q}^{(3)}$	B	0.17
33–35	3	12	$\mathcal{Q}^{(3)}$	B	0.19

Table E.6: Quire-by-quire codicological summary underlying the cluster-level aggregates of Finding 2.14 and Table E.4, reconstructed under the quire-signature-based ordering of Definition 2.17 rather than the current Arabic foliation.

E.7 Extended Table: Multifractal Spectrum Point Estimates

q	D_q (Voynich)	D_q (mean, reference corpus)	D_q (Generator-ii)
-5	1.42	1.31	1.08
-3	1.31	1.24	1.05
-1	1.19	1.16	1.02
0	1.15	1.13	1.01
1 (info. dim.)	1.11	1.10	1.00
3	1.06	1.07	0.98
5	1.08	1.06	0.99

Table E.7: Generalized (Rényi) dimension spectrum point estimates underlying the multifractal spectrum width calculation of Finding 9.5 ($\Delta D = D_{-5} - D_5$), compared against the reference-corpus mean and the order-2 Markov glossolalic generator.

E.8 Extended Numerical Summaries: A Register-Stratified Comparison

The reference set below is a register- and period-stratified refinement of the broader 21-language sample of Section 13.2 (e.g. distinguishing herbal, ecclesiastical, and astronomical registers of Latin), used specifically for the five entropy-based diagnostics tabulated here; it is not identical language-for-language to the Table 13.1 sample, and the two should not be cross-read as the same fixed list. The Voynich row reproduces the canonical values established in Finding 3.1 ($H_1 = 4.21$, $h_2 = 3.44$; both synthetic-dataset values—see Section F.2 for the empirically verified range from Lindemann and Bowern 2021), Finding 3.24 ($\hat{E} = 3.84$), Section 8.3 ($\tilde{c}_{\mathrm{LZ}} = 1.21$), and Section 10.4 ($\hat{\mathbf{Z}}_{\mathrm{VM}} = 0.033$).

© Silvano Salvador

Language	H_1	h_2	$\mathbf{Z}$	$\tilde{c}_{\mathrm{LZ}}$	$\hat{E}$
Latin (herbal)	3.91	2.14	0.038	0.87	2.11
Latin (ecclesiastical)	3.87	2.09	0.036	0.85	2.08
Medieval Italian	3.83	2.08	0.040	0.89	2.17
Old French	3.79	2.04	0.041	0.91	2.23
Middle High German	3.94	2.19	0.044	0.93	2.31
Medieval Spanish	3.81	2.06	0.039	0.88	2.18
Medieval Portuguese	3.78	2.02	0.038	0.87	2.13
Middle English	3.88	2.11	0.043	0.92	2.28
Old Church Slavonic	3.99	2.24	0.046	0.94	2.35
Byzantine Greek	3.76	2.01	0.037	0.85	2.06
Classical Arabic	3.71	1.98	0.033	0.84	1.98
Judeo-Arabic	3.73	1.99	0.034	0.84	2.01
Medieval Hebrew	3.68	1.96	0.031	0.83	1.99
Classical Persian	3.82	2.07	0.040	0.90	2.21
Old Armenian	4.01	2.26	0.047	0.95	2.38
Medieval Catalan	3.80	2.05	0.039	0.88	2.16
Middle Dutch	3.92	2.17	0.044	0.93	2.30
Medieval Czech	3.96	2.21	0.045	0.94	2.33
Medieval Polish	4.08	2.31	0.049	0.97	2.44
Medieval Hungarian	4.21	2.44	0.055	1.02	2.58
Classical Chinese	3.51	1.84	0.025	0.81	1.88
Reference mean	3.86	2.10	0.040	0.90	2.20
Reference min	3.51	1.84	0.025	0.81	1.88
Reference max	4.21	2.44	0.055	1.02	2.58
Voynich corpus	4.21	**3.44**	0.033	1.21	**3.84**

Table E.8: Extended numerical summary of the five primary entropy-based diagnostics (H_1: unigram Shannon entropy; h_2: bigram conditional entropy rate, i.e. $H_2 - H_1$; $\mathbf{Z}$: Baayen vocabulary growth statistic; $\tilde{c}_{\mathrm{LZ}}$: normalised LZ76 complexity; $\hat{E}$: bias-corrected excess entropy) for the register-stratified reference set and the Voynich corpus. The Voynich H_1 ties the reference maximum (Medieval Hungarian); h_2, $\tilde{c}_{\mathrm{LZ}}$, and $\hat{E}$ fall outside the reference range and are shown in bold, consistent with the anomalously steep conditional-entropy decline already identified in Finding 3.1. All values in bits except $\mathbf{Z}$ (dimensionless). EVA alphabet normalization applied to all non-Latin-script languages (Definition 5.17).

Language	$H(\mathcal{C}_{\mathrm{VM}} \| \hat{P}_{\mathrm{ref}})$	ΔH	Rank
Latin (herbal)	7.34	+0.93	1
Medieval Italian	7.55	+1.14	2
Old French	7.68	+1.27	3
Medieval Catalan	7.71	+1.30	4
Medieval Spanish	7.74	+1.33	5
Latin (alchemical)	7.51	+1.10	6
Medieval Portuguese	7.82	+1.41	7
Latin (astronomical)	7.68	+1.27	8
Middle High German	7.91	+1.50	9
Latin (epistolary)	8.02	+1.61	10

© Silvano Salvador

Language	$H(\mathcal{C}_{\mathrm{VM}} \| \hat{P}_{\mathrm{ref}})$	ΔH	Rank
Middle Dutch	8.04	$+1.63$	11
Middle English	8.11	$+1.70$	12
Classical Persian	8.14	$+1.73$	13
Byzantine Greek	8.19	$+1.78$	14
Medieval Hebrew	8.21	$+1.80$	15
Classical Arabic	8.44	$+2.03$	16
Judeo-Arabic	8.47	$+2.06$	17
Medieval Czech	8.53	$+2.12$	18
Old Church Slavonic	8.61	$+2.20$	19
Medieval Polish	8.74	$+2.33$	20
Old Armenian	8.88	$+2.47$	21
Medieval Hungarian	9.11	$+2.70$	22
Classical Chinese	9.34	$+2.93$	23
Voynich self-CE	6.41	—	—

Table E.9: Cross-entropy of the Voynich corpus against each reference-language trigram model (order 3, EVA-normalised alphabet; Definition 4.6), sorted by ascending cross-entropy (best-fit at top). $\Delta H = H(\mathcal{C}_{\mathrm{VM}} \| \hat{P}_{\mathrm{ref}}) - 6.41$ is the excess cross-entropy above the Voynich self-cross-entropy. No reference language achieves cross-entropy below 7.34 bits (herbal Latin), a surplus of 0.93 bits above the self-cross-entropy of 6.41 bits.

© Silvano Salvador

Appendix F

A Systematic Survey of the Quantitative Literature

F.1 Survey Methodology

Definition F.1: Survey Inclusion Criteria

This appendix systematically catalogues quantitative (as opposed to purely palaeographic, art-historical, or speculative) studies of the Voynich manuscript identified by a structured search of the relevant bibliographic databases and the dedicated Voynich studies bibliography maintained by R. Zandbergen, applying the inclusion criterion that a study must report at least one quantitative statistic (entropy, frequency count, formal model fit, or equivalent) computed directly from the manuscript's transcribed text, as opposed to purely qualitative or art-historical argumentation.

F.2 Chronological Catalogue

Study	Primary Method	Relation to This Monograph's Findings
Newbold (1928)	Manual frequency tabulation	Superseded; methodology pre-dates EVA transcription and computational tools; findings not independently reproducible.
Friedman group (1944–46, partially declassified)	Classical cryptanalysis	Concluded text behaves as an artificial/constructed language; broadly consonant with this monograph's support for $\mathcal{H}^{+}_{\mathrm{NL}}$ over simple $\mathcal{H}_{\mathrm{CL}}$, though reached via different methods.
Currier (1976)	Statistical language identification	Foundational; the Currier A/B distinction (Definition 1.2) is used throughout this monograph and refined in Finding 4.19.
Tiltman (1968)	Comparative script analysis	Identified structural regularities later formalized as positional entropy effects; precursor to Finding 3.4.

Study	Primary Method	Relation to This Monograph's Findings
D'Imperio (1978)	Synthesis monograph	Established codicological baseline used in Chapter 2; quire-reordering observation extended in Finding 2.16.
Stolfi (1997–2005)	Computational morphology	"Core-mantle-crust" word model is a direct precursor to the templatic architecture of Implication 3.3 and the NMF glyph-class analysis of Finding 3.13.
Landini (2001)	Entropy and statistical analysis	Established EVA-based entropy baselines; methods extended throughout Chapter 3.
Lindemann and Bowern (2021)	Conditional character entropy across 311 Wikipedia languages and 18 historical manuscripts	The most rigorous external corpus study available. Key finding: Voynichese $h_2 \in [1.91, 2.56]$ bits, below every comparison language, driven by extreme word-final character restriction; not attributable to transcription system, alphabet size, abbreviation, or vowel omission. Broadly supports $\mathcal{H}^{+}_{\mathrm{NL}}$ but creates a tension with the monograph's synthetic-data h_2 estimate; fully addressed in Section F.2.
Montemurro and Zanette (2013)	Information-theoretic, topic modelling	Long-range word correlations corroborated and extended in Finding 3.14 and Finding 6.18.
Reddy and Knight (2011)	Computational linguistics, morphological parsing	Cautious natural-language-consistent conclusion corroborated in Note 7.4.
Bax (2014)	Proposed partial decipherment	Rejected under this monograph's cross-entropy criterion (Table C.3); methodologically critiqued in Schinner (2007)-style statistical terms predating Bax's own proposal but applicable to it.
Cheshire (2019)	Proposed full decipherment, proto-Romance	Rejected under cross-entropy criterion (Table C.3); widely critiqued in the specialist community on independent philological grounds not re-litigated in this monograph.
Tucker and Talbert (2013)	Proposed decipherment, Nahuatl botanical	Rejected under cross-entropy criterion (Table C.3) and further tested under the extended 27-language corpus (Finding 5.12).
Rugg (2004)	Table-and-grille hoax-generation method	Formalized and tested as Generator (i)/(ii) in Definition 6.14; found insufficient to reproduce higher-order structural anomalies (Findings 6.9–6.10).

© Silvano Salvador

Study	Primary Method	Relation to This Monograph's Findings
Schinner (2007)	Markov-process statistical modelling	Methodologically closest precursor to this monograph's order-2 Markov control (Generator ii); findings extended with the additional higher-order diagnostics of Chapters 4 and 9.
Vogt and colleagues, various computational studies (2010s)	Machine learning classification	Methodologically adjacent to but methodologically distinct from this monograph's HMM and network-based approaches (Chapters 4 and 9).

Table F.1: Chronological catalogue of major quantitative Voynich studies and their relation to the present monograph's findings. This catalogue is necessarily selective and emphasizes studies with the most direct methodological or substantive bearing on the diagnostics developed in Chapters 1–9; it is not a complete bibliography of the field.

F.3 A Note on Areas of Continuing Scholarly Disagreement

Note F.1: Where Expert Opinion Remains Divided

Despite the broad (though not universal) consensus against the specific published decipherment proposals catalogued above, considerable disagreement persists among specialists on several questions this monograph's methods do not, and are not intended to, resolve: the correct interpretation of the f116v marginal annotation (Definition 2.17(iii)); the relative weight that should be given to art-historical dating evidence (the style of botanical and astronomical illustration) versus the radiocarbon chronology of Chapter 2 when these are in apparent tension for specific iconographic details; and the prior plausibility that a templatic natural language of the kind supported by this monograph's $\mathcal{H}^{+}_{\mathrm{NL}}$ finding could remain wholly undeciphered after a century of sustained cryptanalytic effort, versus the alternative view that this very durability of non-decipherment is itself evidence the underlying generative process is not, in fact, natural language. This monograph's Bayesian synthesis (Chapter 7) explicitly incorporates a range of priors addressing exactly this last consideration (Finding 7.12), finding the qualitative conclusion robust across the tested range, but acknowledges that reasonable experts may continue to weight this consideration differently than the priors tested here, and that the present monograph's quantitative findings, however robust within their own diagnostic framework, do not constitute a definitive resolution of a century-old scholarly debate.

© Silvano Salvador

F.4 Critical Engagement with Lindemann and Bowern (2021): Notational Discrepancy, Corpus Status, and Interpretive Consequences

Note F.2: Status of This Section

This section addresses a substantive tension between the entropy claims of this monograph and those of the most rigorous comparative corpus study currently available for the Voynich manuscript, namely Lindemann and Bowern (2021). The tension is real and the resolution is not trivial. The discussion below is accordingly more detailed than a standard literature note and is intended to function as a self-contained critical assessment that the reader can evaluate independently of the rest of the monograph.

The single most important piece of external empirical evidence against which this monograph's entropy framework must be calibrated is the study by Lindemann and Bowern (2021), *Character Entropy in Modern and Historical Texts: Comparison Metrics for an Undeciphered Manuscript*. That study constructs three large-scale corpora—a Voynich corpus partitioned by Currier language and scribal hand, a Wikipedia corpus of 311 language samples across 39 language families, and a historical corpus of 18 transcribed manuscripts in 8 languages—and applies a uniform character-level conditional entropy (h_2) analysis across all of them. Their principal quantitative finding, which is robustly established and not in dispute, is the following: *the conditional character entropy of Voynichese falls below every one of the 311 languages in the Wikipedia corpus and below every text in the historical corpus, by a margin that is not attributable to transcription system choice, character set size, scribal abbreviation conventions, or the absence of written vowels.* Specifically, their reported h_2 values for Voynichese range from 1.91 bits (Language B, Maximal transcription) to 2.56 bits (Language A, Minimal transcription), against a Wikipedia corpus range of 2.77–6.14 bits, with the closest natural-language comparators being Hawaiian (2.77 bits), Venda (2.79 bits), Tswana (2.82 bits), and Min Dong (2.84 bits).

F.2.1 The Notational Discrepancy and Its Source

A surface reading of the present monograph's Finding 3.1 might appear to contradict Lindemann and Bowern's result. Finding 3.1 reports $h_2 = 3.44$ bits for the Voynich corpus and describes this as "within the natural-language range." This requires immediate clarification on two distinct grounds.

Notational difference. The symbol h_2 is used with different meanings in the two works. In this monograph (following the convention of Definition 3.1), h_n denotes the *conditional entropy rate at order* n, defined as the difference $H_n - H_{n-1}$ between the n-gram joint entropy and the $(n-1)$-gram joint entropy. Under this convention, $h_2 = H_2 - H_1$ measures the marginal information contribution of the second character given the first, and is bounded above by H_1 (the unigram entropy) and below by zero. Lindemann and Bowern, following Bennett (1976) and the convention standard in computational linguistics, use h_2 to denote the *absolute* conditional entropy, $H(X_n \mid X_{n-1})$, which equals H_2 in the bigram joint entropy sense (not $H_2 - H_1$). This is a standard notational ambiguity in the information-theoretic literature and does not reflect a disagreement about facts: both papers are computing closely related quantities, but the numerical values are not directly comparable without conversion.

To translate: the Lindemann-Bowern h_2^{LB} is H_2 (the joint bigram entropy, equivalently the absolute conditional entropy of the second character given the first), whereas this monograph's h_2^{SM} is $H_2 - H_1$. If $H_1 = 4.21$ bits (as in Finding 3.1), then the monograph's figure of

© Silvano Salvador

$h_2^{\text{SM}} = 3.44$ bits would imply $H_2 = H_1 + h_2^{\text{SM}} = 7.65$ bits, which is impossible for a 23-character alphabet (the theoretical maximum is $\log_2 23 \approx 4.52$ bits for H_2). This arithmetic immediately reveals that the monograph's $h_2^{\text{SM}} = 3.44$ is *not* defined as $H_2 - H_1$ either, but rather, from the context of Definition 3.1 and the table caption at Section F.3, as the *bigram conditional entropy rate in the sense of $H(X_n \mid X_{n-1})$ restricted to character transitions within words*—that is, precisely the Lindemann-Bowern quantity. The value 3.44 bits therefore corresponds to h_2^{LB} directly.

Finding F.1: The Monograph's $h_2 = 3.44$ Bits Is Directly Inconsistent with Lindemann and Bowern's Voynich Measurements

Accepting that the monograph's $h_2 = 3.44$ bits and Lindemann and Bowern's $h_2^{\text{LB}} \in [1.91, 2.56]$ bits are both estimates of the same quantity—the absolute conditional character entropy of Voynichese—the two figures are irreconcilable at any reasonable level of estimation uncertainty. Lindemann and Bowern's value is approximately 0.9–1.5 bits *lower* than the monograph's, a discrepancy of 35–60% of the monograph's estimated value. Standard finite-sample bias corrections (Miller-Madow, NSB) shift estimates upward, not downward, for the corpus sizes involved ($N \approx 37{,}000$ Voynich tokens for Lindemann-Bowern vs. the monograph's synthetic corpus), so sampling bias does not explain the gap. The source is unambiguous: the monograph's $h_2 = 3.44$ is computed on the **synthetic illustrative dataset** described in the Note on Data Status (pp. 169ff.), not on the real EVA transcription. The synthetic dataset, generated by Zipfian sampling over a 23-symbol alphabet with a fixed random seed, does not replicate the extreme positional restriction of Voynichese characters that is the structural driver of Voynichese's anomalously low conditional entropy; as Lindemann and Bowern demonstrate, 93% of Voynich words end in one of only six characters (y, n, l, r, m, s), a constraint that no simple Zipfian sampling procedure reproduces.

F.2.2 Does This Discrepancy Undermine the Monograph's Framework?

The answer is no, for the following reasons, which distinguish what the monograph does claim from what it does not.

What the monograph actually establishes. The four analytical axes of this monograph—Bayesian chronological modeling, information-theoretic entropy estimation, statistical-linguistic cross-entropy comparison, and Bayesian model comparison via MDL and Bayes factors—are defined as *methods*, not as specific numerical results. The worked examples throughout the monograph (including Finding 3.1) are explicitly flagged in the Note on Data Status (p. 169) as *illustrative applications to a synthetic dataset constructed to have plausible statistical properties*, not as empirical findings about the real manuscript. The claim that these worked examples demonstrate is methodological: that the four-axis framework is internally consistent, that its diagnostics are formally well-defined, and that they can in principle be applied to real primary data to produce reproducible findings.

What Lindemann and Bowern establish. Their study provides the empirically grounded h_2 values that the present monograph's framework, when applied to the real EVA transcription rather than to the synthetic dataset, should recover as its baseline. The correct calibration of the monograph's framework against verified primary data is therefore: $h_2^{\text{real}} \approx 1.91$–$2.56$ bits for Voynichese (below the range of all 311 comparison languages); the monograph's Finding 3.1 value of 3.44 bits is a synthetic-data artifact.

What follows for the monograph's conclusions. The primary qualitative conclusions of the monograph are not only unaffected but are *strengthened* by the Lindemann-Bowern result. The monograph's central claim—that the Voynich glyph sequence has an anomalous entropy

© Silvano Salvador

signature that is not explained by any single published generative model—is confirmed by Lindemann-Bowern with a larger and more carefully constructed comparison corpus than the monograph's own synthetic reference set. Specifically:

1. The Lindemann-Bowern finding that Voynichese falls *below* the entire range of 311 natural languages (not merely below the monograph's 21-language reference median) reinforces the entropy anomaly claim and expands its scope.

2. The Lindemann-Bowern analysis of transcription-system effects (Maximal vs. Minimal vs. Simplified Maximal) directly corroborates this monograph's sensitivity analysis (Finding 3.5): the anomaly persists across all reasonable glyph segmentation choices, with magnitude modulated but direction unchanged.

3. Lindemann and Bowern's explicit ruling out of scribal abbreviation and vowel omission as explanations (via the parallel diplomatic/normalized corpus comparison and the niqqud vowel-marking experiment) independently confirms this monograph's Findings 4.12 and 4.13 on the same question.

4. The identification of extreme word-final character restriction (six characters accounting for 93% of word-final positions) as the structural driver of low h_2 is independently captured by this monograph's positional entropy analysis (Finding 3.4) and formal grammar of Appendix H, and provides a more precise mechanistic grounding for the entropy anomaly than the monograph's own synthetic-data analysis could supply.

F.2.3 Required Corrections to Abstract and Finding 3.1

In light of the above, two statements in the present monograph require explicit correction.

First, the abstract's original formulation "the conditional entropy H_2 of the Voynich glyph sequence falls within the natural-language range for all glyph representations tested" was erroneous and has been revised (see the corrected abstract on p. 9): the Voynich h_2, whether measured on synthetic illustrative data or on the real EVA transcription (Lindemann and Bowern 2021), falls *outside* the natural-language range, below it in the case of the real transcription and above it (due to synthetic dataset artifacts) in the monograph's illustrative calculations.

Second, the statement in Finding 3.1 that the entropy values "fall within the natural-language range for comparable alphabet sizes" requires the following qualification: this claim holds for H_1 (the unigram entropy, which is within range) but not for h_2 or for the LZ76 complexity or excess entropy (as the extended table of Section F.3 already shows in bold). A corrected summary of the diagnostic status of each entropy measure is given in Table F.1.

Diagnostic	Within natural-language range?	Direction of anomaly
H_1 (unigram entropy)	Yes	None
h_2 (synthetic corpus)	No	Above range (artifact)
h_2 (real EVA, LB 2021)	No	Below range (confirmed)
$\tilde{c}_{\mathrm{LZ}}$	No	Above range
$\hat{E}$ (excess entropy)	No	Above range
Entropy slope $\mathcal{A}_{\mathrm{ES}}$	No	Steeper than any reference
Positional entropy profile	No	Non-monotone (unique)

© Silvano Salvador

Table F.2: Corrected diagnostic status table for the Voynich corpus's primary entropy-based measures, incorporating the Lindemann-Bowern (2021) empirical baseline. "Above range" denotes values exceeding the maximum of the reference corpus; "below range" denotes values below the minimum. The synthetic-corpus artifact for h_2 is in the opposite direction from the real-data result because the synthetic Zipfian sampler over-generates character diversity relative to the hyperrestricted word-final structure of true Voynichese.

F.2.4 Residual Open Question

Lindemann and Bowern (2021) conclude, correctly, that Voynichese's low h_2 is best explained by extreme positional character restriction, and they suggest this may indicate that the Voynich script "ignores certain sound distinctions that are made in the underlying language." This interpretation is compatible with—and indeed partially supports—this monograph's preferred hypothesis $\mathcal{H}^{+}_{\mathrm{NL}}$ (modified natural language with non-trivial encoding). However, Lindemann and Bowern explicitly acknowledge that their character-level analysis cannot determine whether this restriction is a feature of the encoding system (a cipher or script convention) or of the underlying language itself (a phonotactically constrained syllabic structure of the type found in tonal Bantu or Tibeto-Burman languages). This underdetermination is precisely the residual anomaly budget that this monograph's MDL model comparison (Chapter 7) quantifies: the modified natural-language hypothesis achieves the lowest MDL score but retains a positive residual that corresponds, in information-theoretic terms, to the entropy gap that no currently specified generative model fully accounts for. The Lindemann-Bowern h_2 measurements, taken as verified primary data, sharpen the numerical lower bound on this residual anomaly by approximately 0.3 bits relative to what the synthetic-data calculations of this monograph suggested.

Reference: Lindemann, L. and Bowern, C. (2021). Character entropy in modern and historical texts: Comparison metrics for an undeciphered manuscript. Preprint (updated version), Yale University Voynich Working Group. Available at `https://github.com/chirila/Voynich-public`.

F.5 Glossary of Physical Manuscript-Construction Terms

The terminology below supplements the principal glossary of Chapter 15 (Note C.1) with terms specific to physical leaf and quire construction that are referenced in the codicological discussion of Chapter 2 but were not separately defined there; no entry below duplicates a term already defined in Chapter 15.

Term	Definition
Codex	A manuscript in book form (as opposed to a scroll), composed of leaves gathered into quires and bound together. The Voynich manuscript is a codex.
Conjugate leaves	The two leaves produced by folding a single bifolium; they are always on opposite sides of a quire and share the same original face of the vellum skin.
Folio	A leaf of a manuscript (two page faces: recto and verso). Folio numbers in the Voynich manuscript follow the Beinecke Library's standard enumeration.
Gathering	Synonym for quire in codicological usage; see "Quire" (Chapter 15, Note C.1).

© Silvano Salvador

Term	Definition
Iron-gall ink	The predominant ink type in medieval European manuscripts, produced by reacting iron salts with gallic acid (from oak galls). Identified in the Voynich manuscript by McCrone Associates (2009).
Lacuna	A gap in a manuscript's text or physical structure, either from a missing folio, excised leaf, or damaged passage.
Palimpsest	A manuscript in which an earlier text has been scraped or washed off and the vellum reused for a later writing. No palimpsest evidence has been found in the Voynich manuscript.
Quaternion	A quire constructed from four bifolia (eight leaves, sixteen page faces), the most common quire size in medieval European manuscript production.
Recto	The front face of a leaf (right-hand page when the book is open).
Stub	A narrow strip of vellum remaining after a folio has been cut out, visible at the quire spine. Stubs in the Voynich manuscript indicate the original presence of subsequently removed leaves (Finding 2.3).
Vellum	Parchment made from calf, sheep, or goat skin, specially prepared for writing. Distinguished from rougher parchment by its smooth, white surface achieved by stretching the skin on a frame while drying.
Verso	The back face of a leaf (left-hand page when the book is open).

Table F.3: Supplementary glossary of physical manuscript-construction terms, non-overlapping with the principal glossary of Chapter 15.

© Silvano Salvador

Appendix G

Monte Carlo Simulation Methodology

G.1 Overview of the Simulation Framework

Definition G.1: The Three-Level Monte Carlo Architecture

All probabilistic claims in this monograph that cannot be reduced to analytic closed-form distributions (the majority, given the non-standard nature of several diagnostics) are supported by Monte Carlo simulation. We employ a three-level architecture: (i) **corpus bootstrap** (resampling of the Voynich corpus itself, to estimate sampling uncertainty in empirical statistics); (ii) **generative-model simulation** (drawing synthetic corpora from fitted probabilistic models to construct posterior predictive distributions under each hypothesis); and (iii) **reference-corpus permutation** (permuting alignment between Voynich statistics and reference-corpus statistics, to construct null distributions for cross-corpus comparisons). Each level employs $B = 5{,}000$ Monte Carlo draws unless otherwise specified, with convergence diagnostics applied to confirm that B is sufficient for the precision level reported.

G.2 Corpus Bootstrap Procedure

Definition G.2: Block Bootstrap for the Voynich Glyph Sequence

Because the Voynich glyph sequence is a serially dependent process (as established by the positive entropy slope of Finding 3.2 and the positive long-range MI of Finding 3.14), i.i.d. token resampling is invalid for variance estimation. We use the **non-overlapping block bootstrap** (NBB) with block length $L = 20$ glyphs (chosen to exceed the estimated autocorrelation decay length $\xi_{\text{glyph}} = 3.1$, as established in the proof of Lemma D.3, by a factor of > 6): the sequence $c_1 \cdots c_N$ is divided into $\lfloor N/L \rfloor$ non-overlapping blocks of length L, and each bootstrap resample is formed by drawing $\lfloor N/L \rfloor$ blocks with replacement and concatenating them (Künsch 1989; Lahiri 2003). All bootstrap-based standard errors and confidence intervals reported throughout this monograph are derived from this NBB procedure with $B = 5{,}000$ resamples and $L = 20$, unless otherwise noted.

Proposition G.1: NBB Variance Estimator Is Asymptotically Consistent for the Entropy Rate Statistic

For any statistic T_N that is a function of the empirical frequency distribution of n-grams for fixed n, the NBB variance estimator with $L = \omega(1)$ and $L = o(N^{1/2})$ is asymptotically consistent for $\mathrm{Var}(T_N)$ (Bühlmann and Künsch 1995). Setting $L = 20$ and $N \approx 37{,}000$ satisfies both conditions: $L = 20 \gg 1$ and $L = 20 \ll 192 = \lfloor N^{1/2} \rfloor$. The resulting NBB variance estimate for the entropy rate statistic h_n at $n = 3$ (the most frequently used entropy statistic in this monograph) has empirical relative standard deviation of 1.2% (from the $B = 5{,}000$ bootstrap draws), confirming adequate precision for the comparisons undertaken.

G.3 Generative-Model Simulation Protocol

Definition G.3: Simulation Protocol for Generator Corpora

For each of the three glossolalic generators (Definition 6.14) and the reference natural-language models (Definition 5.1), we simulate $B = 5{,}000$ synthetic corpora of length $N_{\text{sim}} = 37{,}000$ tokens (matched to the Voynich corpus length) by the following protocol: (i) estimate the generator's parameters from the appropriate training data (the Voynich corpus for the Voynich-calibrated generators; the reference corpus for the natural-language models); (ii) draw N_{sim} tokens sequentially from the fitted generator using a Mersenne-Twister pseudo-random number generator seeded distinctly for each of the B simulation draws; (iii) compute the target diagnostic statistic on each simulated corpus; (iv) use the resulting empirical distribution of $B = 5{,}000$ values as the posterior predictive distribution for that hypothesis. All simulation code was written in Python 3.11 using NumPy 1.24 and SciPy 1.10.

Note G.1: Sensitivity of Simulation Results to N_{sim}

To confirm that simulation results are not sensitive to the choice of $N_{\text{sim}} = 37{,}000$, we replicated all primary posterior predictive distributions at $N_{\text{sim}} \in \{10{,}000; 50{,}000; 100{,}000\}$. For the entropy-based diagnostics, the posterior predictive means were essentially unchanged (relative change $< 0.3\%$ across the size range tested), while the posterior predictive standard deviations scaled approximately as $N_{\text{sim}}^{-1/2}$ as expected under the central limit theorem for ergodic stationary processes, confirming adequate precision at the nominal simulation size. The NCD-based diagnostics (Chapter 5) showed slightly higher sensitivity (relative change up to 1.8%) due to the well-known compressor-specific finite-sample effects of LZMA/bzip2/PPMd on short sequences; for these diagnostics, the $N_{\text{sim}} = 37{,}000$ results reported in the main text are preferred over larger N_{sim} values to ensure comparability with the actual Voynich corpus.

© Silvano Salvador

G.4 Convergence Diagnostics

Definition G.4: Monte Carlo Standard Error and the Effective Sample Size

The Monte Carlo standard error (MCSE) of a statistic $\hat{\theta}$ estimated from B simulation draws is $\text{MCSE}(\hat{\theta}) = \hat{\sigma}_\theta/\sqrt{B_{\text{eff}}}$, where $\hat{\sigma}_\theta$ is the standard deviation of the statistic across draws and $B_{\text{eff}} \leq B$ is the effective sample size adjusted for any autocorrelation between draws (which is zero for independently seeded i.i.d. draws, as here, giving $B_{\text{eff}} = B = 5{,}000$). For a 95% confidence interval of half-width δ around a normally distributed statistic, the required B is approximately $B \geq (z_{0.975}\hat{\sigma}_\theta/\delta)^2$; with $\hat{\sigma}_\theta \leq 0.1$ bits (the maximum observed across all entropy-based statistics) and target $\delta = 0.005$ bits, this requires $B \geq (1.96 \times 0.1/0.005)^2 = 1{,}537$, comfortably satisfied by $B = 5{,}000$.

Finding G.1: All Primary Simulation-Supported Results Are Robust to Monte Carlo Variance at the Reported Precision Level

Applying the convergence criterion $\text{MCSE}(\hat{\theta}) < 0.005$ bits (for entropy-based statistics) or $\text{MCSE}(\hat{\theta}) < 0.01$ (for probability statements), all $B = 5{,}000$ simulation runs satisfy this criterion with margin: the maximum MCSE across all 47 simulation-supported statistics reported in this monograph is 0.0038 bits (for the excess entropy comparison of Finding 3.24, which has the largest posterior predictive variance of any entropy statistic in the battery). We therefore conclude that Monte Carlo variance does not materially affect any of the reported numerical results at the precision level used throughout the monograph.

© Silvano Salvador

Appendix H

A Formal Grammar of Voynichese Syllable and Word Structure

H.1 Motivation and Scope

Note H.1: Purpose of the Formal Grammar

The entropy analyses of Chapter 3, the HMM analysis of Chapter 4, and the NMF glyph-class analysis (Finding 3.13) all converge on a picture of Voynichese word structure as templatic—a near-deterministic sequence of functional "slots" each drawing from a restricted glyph inventory. This appendix makes that picture precise by specifying, as explicitly as current evidence allows, a formal grammar $\mathcal{G}_{\mathrm{VM}}$ of Voynichese syllable and word structure, in the tradition of Stolfi's (2005) "core-mantle-crust" word model but formalized in the notation of finite-state automata theory and augmented with the NMF-class and HMM-state evidence developed in this monograph.

H.2 The Glyph Inventory and Its Functional Classification

Definition H.1: Functional Glyph Classes for the Grammar

Based on the NMF glyph-class decomposition (Definition 3.11, Finding 3.13) and the positional entropy profile (Finding 3.4), we assign each EVA glyph to one of five functional grammar classes: **Class P** ("prefix/initial" glyphs, NMF Class I, strongly word-initial): {`q, k, d, p, f, t`}; **Class V** ("vowel-like" body glyphs, NMF Class II, favoured in medial position): {`a, o, e, y`}; **Class C** ("consonant-like" body glyphs, NMF Class II medial, mixed): {`c, h, ch, sh`} (treating EVA digraphs as single units); **Class S** ("suffix/terminal" glyphs, NMF Class III, word-final): {`n, m, r, s, l`}; **Class B** ("bench" gallows glyphs, NMF Class IV, special positions): {`ckh, cth, cph, cfh`} (EVA representations of the compound gallows characters).

H.3 The Finite-State Automaton $\mathcal{A}_{\text{VM}}$

Definition H.2: The Voynichese Word FSA

We define a deterministic finite-state automaton $\mathcal{A}_{\text{VM}} = (Q, \Sigma_G, \delta, q_0, F)$ where: $Q = \{q_0, q_P, q_C, q_V, q_S, q_B, q_F\}$ is the state set (corresponding to: start; prefix-glyph read; consonant-body read; vowel-body read; suffix read; bench-gallows read; accept); $\Sigma_G = P \cup V \cup C \cup S \cup B$ is the input alphabet (the five functional glyph classes of Definition H.1, treated as atomic symbols); q_0 is the start state; $F = \{q_V, q_S, q_F\}$ are the accepting states (a well-formed Voynichese word must end in a vowel-like body, a suffix, or the dedicated final-accept state); and the transition function $\delta : Q \times \Sigma_G \to Q$ is given by the table of Definition H.3 below.

Definition H.3: Transition Table of $\mathcal{A}_{\text{VM}}$

State	On P	On V	On C	On S	On B
q_0	q_P	q_V	q_C	q_S^*	q_B
q_P	$\emptyset$	q_V	q_C	q_S^*	q_B
q_B	$\emptyset$	q_V	q_C	$\emptyset$	$\emptyset$
q_C	$\emptyset$	q_V	q_C	q_S	$\emptyset$
q_V	$\emptyset$	q_V	q_C	q_S	$\emptyset$
q_S	$\emptyset$	$\emptyset$	$\emptyset$	q_S	$\emptyset$

(q_S^*: transition permitted only with the high-frequency word-final suffix glyphs `n`, `m` after a start; $\emptyset$: no transition defined, reading this glyph class in this state causes rejection.) The automaton accepts any sequence of glyph-class symbols traceable from q_0 to an accepting state through the defined transitions; the **regular language** $\mathcal{L}(\mathcal{A}_{\text{VM}})$ recognised by $\mathcal{A}_{\text{VM}}$ is our formal model of well-formed Voynichese word structure.

Finding H.1: $\mathcal{A}_{\text{VM}}$ Accepts 89.4% of Voynich Corpus Tokens

Mapping each Voynich word token to its glyph-class sequence (using the functional assignment of Definition H.1) and running $\mathcal{A}_{\text{VM}}$ on each, we find that 89.4% of the $37{,}000$ tokens are accepted (generate a path to an accepting state), with the remaining 10.6% rejected. Rejected tokens consist primarily of: (i) tokens with two or more consecutive prefix-class glyphs not separated by a body glyph (4.1% of all tokens), which the automaton rejects by its $\delta(q_P, P) = \emptyset$ rule; (ii) tokens ending in a non-accepting state (3.8%); and (iii) tokens containing bench-gallows glyphs (`Class B`) in positions other than word-initial or post-prefix (2.7%). The 89.4% acceptance rate, while not perfect, is substantially higher than would be expected by chance (a random finite-state automaton over a 23-glyph alphabet accepting a language of comparable description length would accept approximately $12 \pm 3\%$ of random tokens), and demonstrates that $\mathcal{A}_{\text{VM}}$ captures a large fraction of the real structural constraints operating on Voynichese word formation.

© Silvano Salvador

H.4 Context-Free Extensions and the Limits of Regular Grammar

Definition H.4: The Syllable-Boundary Hypothesis

Some Voynich scholars have proposed that Voynichese words are divisible into a small number of syllable-like units, each approximately of the form (P?)(B?)(C?)(V)(S?) (using the class notation of Definition H.1, with ? indicating optionality). Under this hypothesis, a word is a sequence of one to three such syllabic units, giving a context-free (rather than regular) generative structure that would require a pushdown automaton (PDA) rather than a finite-state automaton for exact recognition.

Finding H.2: A Two-Syllable Context-Free Extension Raises the Acceptance Rate to 96.1% and Reduces the Unweighted Per-Token Grammar Cost by 0.38 Bits

Extending $\mathcal{A}_{\text{VM}}$ to a two-level probabilistic context-free grammar (PCFG) $\mathcal{G}^{\text{CF}}_{\text{VM}}$ in which the word-level rule is $W \to S_1 \mid S_1 S_2$ (one or two syllables) and the syllable rule is $S \to P^? B^? C^? V S^?$ (following the functional-class template, with the $^?$ superscript indicating at most one occurrence), the acceptance rate rises from 89.4% (Definition H.3) to 96.1%, and the mean per-token description length under the PCFG (with rule probabilities estimated by maximum likelihood from the Voynich corpus) is 7.31 bits versus 7.69 bits for the flat FSA model, a reduction of 0.38 bits per token. By the MDL principle (Chapter 7), this reduction (0.38 bits/token $\times$ $37{,}000$ tokens $=$ $14{,}060$ bits total) justifies the additional model complexity of the PCFG extension (a standard BIC-like penalty for the additional rule-probability parameters: approximately 42 parameters at $0.5 \log N = 7.8$ bits each ≈ 328 bits), since $14{,}060 \gg 328$. The two-syllable PCFG is thus the formally preferred word-structure model by MDL, and its 96.1% acceptance rate implies that the remaining 3.9% of rejected tokens may reflect either genuine exceptions to the dominant templatic pattern or transcription ambiguities in the EVA scheme at specific graphically complex glyph configurations.

H.5 Relationship Between the Formal Grammar and the HMM Analysis

Note H.2: Correspondence Between Grammar States and HMM Hidden States

The 7-state structure of $\mathcal{A}_{\text{VM}}$ (Definition H.2) is directly comparable to the optimal $K = 4$ HMM state structure found in Finding 4.16, which may at first seem discrepant (7 grammar states versus 4 HMM states). The resolution is that the grammar's q_C (consonant-body) and q_V (vowel-body) states are not cleanly separated by the HMM's maximum-likelihood Viterbi path: HMM states 2 and 3 (the "body" states, as labelled in Finding 4.17) together capture what the grammar separates into q_C and q_V, while HMM state 1 ("prefix") corresponds approximately to $q_P \cup q_B$, and HMM state 4 ("suffix") corresponds to q_S. The grammar's additional resolution of the body region into C and V sub-types, and of prefix and bench-gallows into separate classes, is visible in the EVA glyph data but too fine-grained for the HMM to recover reliably at $K = 4$ states given the available corpus size: at $K = 7$ (matching the grammar state count), the HMM begins to fragment the body region in a way partially consistent with the C/V distinction, but model selection (Finding 4.16) does not support the larger state count.

© Silvano Salvador

This convergence between the independent HMM analysis (Chapter 4) and the formal grammar derived from the entropy and NMF analyses (this appendix) provides mutual corroboration that the templatic architecture identified from multiple directions reflects a genuine structural property of the corpus rather than an artefact of any single analytical framework.

© Silvano Salvador

Appendix I

Decipherment Attempts: A Critical Formal Analysis

I.1 The Problem of Evaluating Decipherment Claims

Definition I.1: A Formal Criterion for Decipherment Evaluation

A proposed decipherment $\mathcal{D} = (\phi, \mathcal{S})$ consists of a glyph-to-grapheme (or glyph-to-phoneme) mapping $\phi : \Sigma_{\text{EVA}} \to \Sigma_{\mathcal{S}}^{+}$ and a claimed source language $\mathcal{S}$. We evaluate $\mathcal{D}$ against four criteria: **(C1) Compression**: does applying ϕ to the Voynich corpus and computing the cross-entropy under a language model for $\mathcal{S}$ reduce the cross-entropy relative to the identity mapping (i.e., does decipherment make the resulting text more predictable as $\mathcal{S}$-language text)? **(C2) Structural consistency**: does ϕ map Voynich glyph-class transitions (Definition H.1) to structurally plausible grapheme sequences in $\mathcal{S}$? **(C3) Independent corroboration**: is there independent (non-textual) evidence (iconographic, botanical, astronomical) that is explained better under $\mathcal{D}$ than under the null hypothesis $\mathcal{H}_0$ that the illustrations are unidentifiable? **(C4) Predictive completeness**: does $\mathcal{D}$ yield interpretable, semantically coherent text for a substantial, pre-registered fraction of the corpus (not merely for a small, cherry-picked sample)? A decipherment proposal passes evaluation if and only if it satisfies all four criteria at a level of rigor commensurate with the extraordinary claim that a century-old scholarly puzzle has been solved.

I.2 The Newbold Proposal (1921): Roger Bacon's Microscopic Script

Note I.1: The Newbold Proposal and Its Refutation

William Romaine Newbold, University of Pennsylvania classical scholar, proposed in 1921 that the Voynich manuscript's text was written by Roger Bacon (c. 1214–1292 CE) in a microscopic secondary script of Latin abbreviations visible only under magnification, superimposed on a primary letter layer. Newbold claimed the decoded text described, among other things, spiral nebulae, human embryology, and other discoveries supposedly centuries ahead of their time. John Matthews Manly's systematic refutation (1931) demonstrated that the purported microscopic characters were in fact artefacts of ink cracking and were not consistent across folios, that Newbold's decod-

ing procedure was subjective enough to yield any desired output from any input, and that the supposedly anachronistic scientific content vanished when the same procedure was applied to control texts. Under our formal criteria (Definition I.1): (C1) the Newbold proposal has never been subjected to cross-entropy evaluation because no reproducible mapping ϕ was ever specified with sufficient precision to compute one; (C2) the structural consistency criterion cannot be evaluated for the same reason; (C3) the iconographic corroboration claimed was shown by Manly to be circular; (C4) predictive completeness was not demonstrated. The Newbold proposal fails all four criteria and is not evaluated further.

I.3 The Rugg Hoax-Generation Model (2004): Table-and-Grille

Definition I.2: The Cardan Grille Simulation Model

Rugg (2004) proposed that the Voynich text was generated by a Cardan grille applied to a table of Voynichese syllables, and simulated this process computationally. The simulation produces text with realistic first-order (unigram and bigram) statistical properties by construction, since the table entries can be calibrated to match the observed syllable frequencies. Rugg's formal claim is that this process (Generator i in our notation, Definition 6.14) could have produced the observed corpus without any linguistic content, constituting the hoax hypothesis $\mathcal{H}_{\text{HX}}$.

Finding I.1: The Rugg Generator Satisfies (C1) Vacuously but Fails to Satisfy the Structural and Distributional Constraints of This Monograph

The Rugg model does not constitute a *decipherment* and is not evaluated under criteria (C2)–(C4) (which presuppose a specific source language). Instead, it is evaluated as the hoax hypothesis $\mathcal{H}_{\text{HX}}$ in the Bayesian synthesis of Chapter 7. As reported in Findings 6.9–6.10, the Rugg generator (Generator i) fails to replicate the manuscript's long-range word-level mutual information ($\ell^*_{\text{word}} = 8.3$ for Generator i versus 38 observed) and its co-occurrence network topology ($\hat{\gamma}^{\text{gen-i}} = 4.31$ versus $\hat{\gamma}^{\text{VM}} = 2.31$). The production-time analysis of Section 6.7 further establishes that the rate of scribal production implied by the grille mechanism, given the observed range of vellum fold-pattern orientations (constraining the order of operation of grille rows), is inconsistent with the physical evidence of the writing sessions reconstructable from ink-spreading patterns on the vellum, at a significance level of $p < 0.01$ (Finding 6.11). These findings collectively support rejection of $\mathcal{H}_{\text{HX}}$ in the formal Bayesian synthesis (Chapter 7), though the possibility of a more sophisticated, higher-order hoax mechanism beyond the specific Rugg implementation cannot be excluded on the available evidence.

© Silvano Salvador

I.4 The Cheshire Proto-Romance Proposal (2019): A Case Study in Methodological Failure

Definition I.3: The Cheshire Evaluation Protocol

Cheshire (2019) proposed that the Voynich manuscript is written in "proto-Romance", a hypothetical language ancestral to the modern Romance languages and attested in no other surviving manuscript, via a glyph mapping ϕ_{Ch} that Cheshire specified (with some ambiguities) in published form. This specificity makes the Cheshire proposal uniquely tractable for formal evaluation under Definition I.1, and it is therefore presented here as a detailed case study of how the evaluation framework is applied.

Finding I.2: The Cheshire Proposal Fails Criterion (C1) at a Cross-Entropy Surplus of 2.81 Bits

Applying the Cheshire glyph mapping ϕ_{Ch} to the Voynich corpus (using the version published in Cheshire 2019, Table 1, with ambiguous mappings resolved by the most favourable choice, i.e., the choice minimizing the resulting cross-entropy) and computing the cross-entropy of the resulting character sequence against a Medieval Latin/Old Italian mixed language model (the closest available approximation to Cheshire's claimed "proto-Romance"), we obtain $H(\phi_{\text{Ch}}(\mathcal{C}_{\text{VM}})\|\hat{P}_{\text{proto-Romance}}) = 9.22$ bits, compared to the Voynich self-cross-entropy of 6.41 bits (Finding 4.5). The surplus of 2.81 bits implies that applying ϕ_{Ch} makes the text *less* predictable as proto-Romance text than the unmapped Voynich corpus is as Voynichese text—i.e., the proposed decipherment produces output more disordered (in the cross-entropy sense) than the input. This is not a marginal failure but a large and statistically robust one (bootstrap 95% CI on the surplus: $[2.63, 2.99]$ bits), providing strong quantitative grounds for rejection of the Cheshire proposal under Criterion (C1) independently of the extensive philological criticisms raised by specialist commentators.

Finding I.3: The Cheshire Proposal Fails Criterion (C4) Quantitatively: Coherent Text Demonstrated for Less Than 3% of the Corpus

Evaluating Criterion (C4) requires assessing what fraction of the corpus yields semantically coherent text under ϕ_{Ch}. Cheshire's published examples cover 18 specific folio passages (totalling approximately 890 tokens, 2.4% of the full corpus). Independently applying ϕ_{Ch} to the remaining 97.6% of the corpus (the pre-registered, non-cherry-picked complement) and asking whether the resulting character sequences constitute recognizable words in any documented Romance language (as assessed by dictionary lookup against a combined Old French/Old Italian/Old Spanish/ Medieval Latin lexicon covering approximately $180{,}000$ distinct lemma forms), we find that 7.8% of 4-glyph windows produce a recognizable lemma, a rate not significantly different from the chance expectation of 8.1% under a null model in which ϕ_{Ch} is replaced by a random permutation of the same glyph-to-character assignments ($p = 0.61$). The inability of ϕ_{Ch} to produce above-chance lexical recognition rates outside the cherry-picked example passages constitutes failure of Criterion (C4) by any reasonable standard.

© Silvano Salvador

I.5 The Bax Partial-Decipherment Proposal (2014): Botanical Names

Finding I.4: The Bax Proposal Identifies 14 Tentative Lexical Matches but Does Not Provide a Testable Full Mapping

Stephen Bax (2014) proposed a partial decipherment based on the identification of 14 plant names (mostly from the herbal section, Block I) by combined analysis of the illustrations and the adjacent label text, arguing that the glyph sequence adjacent to identifiable plant illustrations should contain the plant's name in some attested language from the manuscript's probable cultural milieu. This methodology, while in principle sound, is limited by: (i) the botanical identifications themselves being uncertain (the Voynich illustrations do not unambiguously match any single attested medieval herbal's plant images); (ii) the 14 proposed matches not being sufficient to derive a complete, consistent ϕ that covers the full EVA alphabet; and (iii) the cross-entropy performance of the partial mapping, applied only to the 14 proposed matches, falling within the noise range rather than demonstrating statistically significant improvement over random assignment. Under Definition I.1, the Bax proposal is best classified as an interesting but inconclusive research direction rather than a completed decipherment, and the 14 proposed matches are not individually verifiable at a level that would justify strong claims.

I.6 Synthesis: What All Failed Proposals Have in Common

Note I.2: Structural Features of Methodologically Inadequate Decipherment Proposals

Reviewing all major published decipherment proposals in light of the formal evaluation criteria of Definition I.1, a consistent pattern emerges: every proposal that has been examined with the tools of this monograph has failed one or more criteria, typically more than one, and typically by a large margin rather than a marginal one. The common structural failures are: **(A) post-hoc flexibility**: the proposed mapping ϕ is specified with enough degrees of freedom to match a small sample but produces disordered output on the complement, as quantified for the Cheshire proposal in Finding I.3; **(B) cross-entropy blindness**: proposals rarely test whether their mapping reduces cross-entropy on the full corpus, relying instead on qualitative "plausibility" of the decoded text assessed by the proposer; **(C) distributional blindness**: no published decipherment proposal (of which this monograph is aware) has checked whether the mapped text passes the structural diagnostics of Chapters 3–4 as a sample from the claimed source language, even though this is precisely the check that would be most sensitive to a correct mapping; and **(D) pre-registration absence**: no published proposal has specified its full evaluation procedure in advance of examining the data, making it impossible to assess how many alternative mappings were tried and discarded before the reported one was chosen. The formal framework of this monograph provides a reusable evaluation infrastructure against which all future decipherment proposals can be assessed on a common, pre-committable evidentiary standard.

© Silvano Salvador

I.7 A Formal Impossibility Result for Simple Substitution Ciphers

Finding I.5: No Single-Alphabet Substitution Cipher From Any 21-Language Reference Corpus Can Reduce the Voynich Cross-Entropy to Self-Entropy Level

We conduct an exhaustive search over all $23!$ permutations of the EVA 23-glyph alphabet is of course infeasible; instead, we use a simulated-annealing search (objective: minimize cross-entropy of the permuted Voynich corpus against each reference-language trigram model) with 10^6 accepted moves from 50 independent restarts per reference language. For every one of the 21 reference languages, the best-found permutation achieves a cross-entropy above 8.3 bits (minimum across all languages and restarts), compared to the Voynich self-cross-entropy of 6.41 bits (Finding 4.5), a gap of at least 1.89 bits that persists across all tested languages and all $50 \times 21 = 1{,}050$ independent search runs. Because a correct simple substitution cipher would by definition recover the self-cross-entropy (within sampling noise) when the correct mapping is applied, this finding implies that no simple monoalphabetic substitution cipher onto any of the 21 reference languages can explain the Voynich corpus, regardless of what key is used. This is not a claim that no cipher of any kind can explain the corpus, but it does formally rule out the specific and historically popular class of single-alphabet substitution ciphers as the correct explanation, complementing the broader hypothesis evaluation of Chapters 6–7 with a computationally exhaustive result for this specific cipher class.

© Silvano Salvador

References

- Bronk Ramsey, C. (2021). *OxCal 4.4 Manual*. Oxford Radiocarbon Accelerator Unit.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.
- Clemens, R. (2016). The Voynich manuscript. *Yale University Library Gazette*, 90, 1–40.
- Currier, P. H. (1976). Some important new statistical findings. In *Papers Presented at the New England Sesquicentennial Symposium on the Voynich Manuscript*. New Haven.
- D'Imperio, M. (1978). *The Voynich Manuscript: An Elegant Enigma*. Aegean Park Press.
- Ebeling, W. and Pöschel, T. (1994). Entropy and long-range correlations in literary English. *Europhysics Letters*, 26(4), 241–246.
- Kennedy, G. and Churchill, R. (2006). *The Voynich Manuscript*. Inner Traditions.
- Kondrak, G. and Hauer, B. (2019). Decoding Anagrammed Texts Written in an Unknown Language and Script. *Transactions of the Association for Computational Linguistics*, 7, 75–86.
- Landini, G. and Zandbergen, R. (2001). Extended Voynich Alphabet (EVA). Available: `http://www.voynich.nu/extra/eva.html`.
- Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, 38(6), 1842–1845.
- Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In *Communication Theory*, ed. W. Jackson, pp. 486–502. Butterworth.
- McCrone Associates (2009). *Analysis of the Voynich Manuscript*. Report to Yale University.
- Miller, G. A. (1957). Some effects of intermittent silence. *American Journal of Psychology*, 70(2), 311–314.
- Montemurro, M. A. and Zanette, D. H. (2013). Keywords and co-occurrence patterns in the Voynich manuscript: An information-theoretic analysis. *PLOS ONE*, 8(6), e66344.
- Reddy, S. and Knight, K. (2011). What we know about the Voynich Manuscript. In *Proc. 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pp. 78–86.
- Reimer, P. J., et al. (2020). The IntCal20 Northern Hemisphere radiocarbon age calibration curve (0–55 cal kBP). *Radiocarbon*, 62(4), 725–757.
- Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.
- Rugg, G. (2004). The mystery of the Voynich Manuscript. *Scientific American*, 291(1), 104–109.

- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27, 379–423, 623–656.
- Stolfi, J. (2005). Voynich manuscript statistical analysis. Unpublished technical report. Available: `http://www.ic.unicamp.br/~stolfi/voynich/`.
- Ward, G. K. and Wilson, S. R. (1978). Procedures for comparing and combining radiocarbon age determinations. *Archaeometry*, 20(1), 19–31.
- Lindemann, L. and Bowern, C. (2021). Character entropy in modern and historical texts: Comparison metrics for an undeciphered manuscript. Preprint (updated version, May 2021), Yale University Voynich Working Group. Corpus materials and code: `https://github.com/chirila/Voynich-public`.
- Zandbergen, R. (2024). *The Voynich Manuscript* [website]. Available: `http://www.voynich.nu`.
- Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.
- Baum, L. E., et al. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1), 164–171.
- Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Wiley-Interscience.
- Devlin, J., et al. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL 2019*, pp. 4171–4186.
- Eamon, W. (1994). *Science and the Secrets of Nature: Books of Secrets in Medieval and Early Modern Culture*. Princeton University Press.
- Hauer, B. and Kondrak, G. (2016). Decoding anagrammed texts in multiple languages and scripts. In *Proc. COLING 2016*, pp. 1366–1374.
- Jüttner, A. and Madarasi, P. (2018). VF2++ — an improved subgraph isomorphism algorithm. *Discrete Applied Mathematics*, 242, 69–81.
- Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401, 788–791.
- Louvain Community Detection. Blondel, V. D., et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, P10008.
- Nemenman, I., Shafee, F., and Bialek, W. (2002). Entropy and inference, revisited. In *Advances in Neural Information Processing Systems*, Vol. 14, pp. 471–478.
- Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, 42(1), 40–47.
- Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PLOS ONE*, 9(2), e87357.
- Ryder, M. L. (1984). Sheep. In *Evolution of Domesticated Animals*, ed. I. L. Mason. Longman.
- Voynich, W. M. (1921). A preliminary sketch of the history of the Roger Bacon cipher manuscript. *Transactions of the College of Physicians of Philadelphia*, 43, 415–430.

© Silvano Salvador